UNIVERSITY OF DELHI

RATAN TATA
LIBRARY

RATAN TATA LIBRARY

Call No. B280X9J

Accession No. 7502

Date of release for loan

This book should be returned on or before the date last stamped below. An overdue charge of 10 Paise will be collected for each day the book is kept overdue.

~~27 AUG 1976~~

~~2 NOV 1976~~

12 OCT 1981

14 NOV 1981

STATISTICAL METHODS

STATISTICAL METHODS

APPLIED TO AGRICULTURAL ECONOMICS

By

FRANK A. PEARSON

and

KENNETH R. BENNETT

Cornell University

NEW YORK: JOHN WILEY & SONS, INC.
LONDON: CHAPMAN & HALL, LIMITED

Copyright, 1942
BY
FRANK A. PEARSON
AND
KENNETH R. BENNETT

All Rights Reserved
This book or any part thereof must not be reproduced in any form without the written permission of the publisher.

THIRD PRINTING, OCTOBER, 1949

Printed in U. S. A.

PREFACE

This book was written primarily for those interested in applications of statistical methods to agricultural economics. The illustrations are largely drawn from the fields of farm management, marketing, and prices. However, these illustrations are similar to those which might have been taken from other fields in agricultural economics and business. The volume is designed for use as a textbook in colleges and universities or as a general reference book for statistical workers.

The arrangement follows the usual procedure: measures of central tendency, variation, association, and reliability.

The book differs from most textbooks in that it contains two chapters on the tabular analysis of relationships. This subject is ignored in most textbooks despite the fact that it has been and will continue to be the most widely used method of analyzing relationships.

In the chapters on testing reliability, emphasis is placed on problems which arise in the social sciences. The application of tests of significance to tabular analysis is given in chapters 18, 20, and 21; and to correlation analysis, in chapter 22.

We are indebted to R. A. Fisher and associates for the development of many of the newer techniques in testing reliability. We are also indebted to C. H. Goulden and G. W. Snedecor for generous permission to reproduce certain tables; and to Mrs. J. V. Cassetta for editing the manuscript. Of course, the full responsibility for inaccuracies rests with us.

FRANK A. PEARSON
KENNETH R. BENNETT

ITHACA, NEW YORK
July, 1941

CONTENTS

CHAPTER PAGE
1. Frequency Distributions 1
2. Measures of Central Tendency 16
3. Dispersion 36
4. Index Numbers 55
5. Secular Trend 76
6. Seasonal Variation 90
7. Cycles 104
8. Tabular Analysis of Relationships 120
9. Correlation 143
10. Multiple Correlation 166
11. Partial Correlation 185
12. Curvilinear Correlation 200
13. Index of Multiple Correlation 212
14. Joint Correlation 246
15. Tabulation *vs.* Correlation Analysis 264
16. Measures of Reliability 300
17. Standard Errors 304
18. Application of Standard Errors to Tabular Analysis 323
19. The Analysis of Variance 345
20. Application of Analysis of Variance to Tabular Analysis 370
21. Chi Square 387
22. Reliability of Correlation Analysis 401
Appendix 421
Index 435

CHAPTER 1

FREQUENCY DISTRIBUTIONS

Because the human mind is unable to grasp facts contained in large, unordered masses of data, it is necessary to rearrange, condense, or simplify these data in some fashion. Incomes on 89 New York fruit farms for 1913 are illustrative of ungrouped and disordered data (table 1, part *a*). With considerable effort, the reader may note that the highest income was $5,205; and the lowest, – $1,407; and that the other 87 incomes fell between these two extremes. When the incomes are arranged

TABLE 1.—UNARRANGED AND ARRANGED DATA

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

(*a*) Unarranged data				(*b*) Arranged according to magnitude			
$ 1,372	$ 1,887	$ 4,127	$ – 467	$ –1,407	$290	$ 965	$1,887
587	387	1,867	5,205	– 587	387	1,080	1,900
403	961	259	2,897	– 573	403	1,108	1,948
1,167	1,471	– 89	866	– 467	416	1,167	1,965
1,965	1,202	535	216	– 328	421	1,171	2,031
1,879	1,900	2,111	40	– 316	498	1,202	2,111
62	421	965	2,620	– 206	535	1,271	2,194
– 206	– 328	255	416	– 194	577	1,272	2,204
1,271	– 46	1,748	2,204	– 186	618	1,348	2,347
1,272	907	2,194	1	– 111	661	1,372	2,390
4,136	2,347	1,171	– 316	– 89	728	1,409	2,544
951	2,544	1,108	1,452	– 85	735	1,425	2,620
2,031	1,425	1,348	2,735	– 75	735	1,444	2,732
– 194	577	– 186	498	– 46	801	1,452	2,735
–1,407	1,948	2,871	2,732	+ 1	819	1,463	2,820
1,523	– 573	854	1,463	+ 9	845	1,471	2,871
290	2,820	855	1,444	17	854	1,523	2,897
3,702	– 111	1,585	845	40	855	1,543	3,702
9	728	2,390	17	62	866	1,585	4,013
801	1,409	– 85	– 75	216	907	1,748	4,127
1,080	4,013	735	618	255	951	1,867	4,136
4,673	819	1,543	661	259	961	1,879	4,673
			735				5,205

according to size, these same facts can be determined at a glance (table 1, part *b*). However, this rearrangement does not materially increase the ease of further analysis of the data.

A frequency distribution groups the items of a series according to their size and shows the frequency of occurrence of each group. The incomes have been grouped into 14 classes (table 2). Each class extends over a range of $500. The number of farms in each class is shown. For instance, there were 17 farms with incomes from + $500 to + $999. The construction of such a table involves: (1) choosing the size of the class and the number of classes, (2) choosing the class limits, and (3) counting the number of items in each class.

TABLE 2.—FREQUENCY DISTRIBUTION WITH $500 CLASS INTERVAL

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Frequency
−1,500 to −1,001	1
−1,000 to − 501	2
− 500 to − 1	11
0 to 499	14
500 to 999	17
1,000 to 1,499	15
1,500 to 1,999	10
2,000 to 2,499	6
2,500 to 2,999	7
3,000 to 3,499	0
3,500 to 3,999	1
4,000 to 4,499	3
4,500 to 4,999	1
5,000 to 5,499	1
Total	89

NUMBER OF CLASSES

In general, a frequency table[1] should not contain fewer than 8 to 10 classes or more than 25 to 30, depending upon the total number of items in the series.

A series containing a large number of items can be divided into more classes than a series with a small number, because it can supply a considerable frequency to more classes and because random fluctuations among frequencies tend to iron out as the number of items increases. The most desirable frequency table is the one which gives the reader the most information in clearest fashion. Most readers can grasp ideas more readily when only a few classes are used. On the other hand, some of the characteristics of a distribution tend to be obscured with an insufficient number of classes. Frequency distributions with a large number of classes are likely to contain all the characteristics of the series, but it is usually difficult for the reader to ascertain these characteristics.

---

[1] Although frequency distributions are usually shown in tabular form, they may also be shown graphically. However, in this chapter, the terms "frequency distribution" and "frequency table" are often used synonymously.

The calculation of statistical measures from ungrouped data is very difficult. This difficulty is overcome in large part when the data are grouped. The work of calculation from frequency tables is about proportionate to the number of classes.

TABLE 3.—FREQUENCY DISTRIBUTION WITH $2,000 AND $250 CLASS INTERVALS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

$2,000 Classes		$250 Classes	
Class interval, dollars	Frequency	Class interval, dollars	Frequency
−2,000 to − 1	14	−1,500 to −1,251	1
0 to 1,999	56	−1,250 to −1,001	0
2,000 to 3,999	14	−1,000 to − 751	0
4,000 to 5,999	5	− 750 to − 501	2
Total	89	− 500 to − 251	3
		− 250 to − 1	8
		0 to 249	6
		250 to 499	8
		500 to 749	7
		750 to 999	10
		1,000 to 1,249	5
		1,250 to 1,499	10
		1,500 to 1,749	4
		1,750 to 1,999	6
		2,000 to 2,249	4
		2,250 to 2,499	2
		2,500 to 2,749	4
		2,750 to 2,999	3
		3,000 to 3,249	0
		3,250 to 3,499	0
		3,500 to 3,749	1
		3,750 to 3,999	0
		4,000 to 4,249	3
		4,250 to 4,499	0
		4,500 to 4,749	1
		4,750 to 4,999	0
		5,000 to 5,249	1
		Total	89

The labor incomes on 89 farms were grouped into $500 class intervals (table 2). The number of classes, 14, was sufficient to show the main characteristics of the series without confusing the reader with too much detail. When this series was grouped by $2,000 class intervals, 56 farms, or almost two-thirds of the total, were included in the class $0 to $1,999 (table 3). This frequency table clearly shows the reader that the most common labor income on these fruit farms was between $0 and + $1,999. It does not tell the reader the relative proportion of farmers receiving a low income $0 to + $499, or a high income, $1,500 to $1,999. A further division—that is, more classes—would be desirable in this case, both for the reader's information and for further statistical analysis.

When this series was divided into $250 class intervals, the number of classes was 27 (table 3). This frequency distribution contains the important characteristics of the series of labor incomes, but the reader has considerable difficulty in grasping them because of the lack of concentration. The original series does not contain enough observations to support 27 classes. The high variability in the frequencies of contiguous classes is an indication of too many classes for the size of the series; or conversely, not a large enough series for the number of classes. This grouping into 27 classes is little improvement upon the array in informing the reader; neither is it a great improvement in facilitating the calculation of further statistical measures.

SIZE OF CLASS INTERVAL

Strictly speaking, the size of the class interval is determined by the number of classes and the total range in the data. There are, however, certain additional points which one should consider in choosing the size of the classes. The size of the interval should not be such that it tends to obscure or distort the characteristics of the series. If there is no danger of this difficulty, class intervals should be of such common sizes as 2, 5, 10, 25, 50, 100, 500, 1,000, and so on, rather than 1.5, 6, 11, 23, 53, 97, 472, and the like. The human mind is accustomed to thinking in terms of certain multiples of 2, 5, 10, and the like. In the frequency distribution of incomes in table 2, the size of the class intervals was $500, rather than $450 or $535, because 500 is a number that is easy to manipulate mentally and mechanically, and yet does not disturb the major characteristics of the series.

In general, class intervals should be equal in size. A justification for unequal class intervals lies in the saving of space on the printed page.

There is often difficulty in making frequency distributions of size of farms in some sections of the United States because of the tendency for farms to contain 80, 160, or other multiples of 40 and 80 acres. The 160 Illinois farms were grouped by the 50-acre classes 20–69, 70–119, etc. (table 4). There were 52 farms in the class 70–119 acres, 37 of which were exactly 80 acres in size. As a result, the actual average of the class was 83, while the midpoint was 95. The next higher class, 120–169 acres, contained 33 farms of exactly 160 acres. The actual average of this class was 151, six acres above the midpoint, because 160 was within 10 acres of the upper limit of the class. When these farms were grouped into 40-acre classes, 20–59, 60–99, etc., the midpoint was usually the most common acreage and checked quite closely with the average of the class (table 4).

In the calculation of statistical measures from the series of data, the class intervals of 40 acres would give more reliable results than the 50-acre class intervals.

TABLE 4.—EFFECT OF SIZE OF CLASS INTERVALS ON REPRESENTATIVENESS OF FREQUENCY DISTRIBUTIONS

SIZE OF 160 FARMS, BUREAU COUNTY, ILLINOIS, 1930

50-acre intervals				40-acre intervals			
Class interval, acres	Frequency	Actual class average	Midpoint of class	Class interval, acres	Frequency	Actual class average	Midpoint of class
20– 69	15	47	45	20– 59	11	41	40
70–119	52	83	95	60– 99	50	77	80
120–169	58	151	145	100–139	19	117	120
170–219	11	184	195	140–179	50	160	160
220–269	15	240	245	180–219	6	202	200
270–319	3	277	295	220–259	15	240	240
320–369	3	333	345	260–299	3	277	280
370–419	2	402	395	300–339	2	320	320
420–469	0	—	445	340–379	1	360	360
470–519	0	—	495	380–419	2	402	400
520–569	1	560	545	420–459	0	—	440
				460–499	0	—	480
				500–539	0	—	520
				540–579	1	560	560
Total	160	—	—	Total	160	—	—

LOCATION OF CLASS LIMITS

The limits of the classes should be such that the characteristics of the series are not obscured or distorted. The frequency table first of all must tell the truth. So far as possible, there should be symmetrical distribution of the items within each class. The class limits should be chosen so that the midpoint of the class is representative of all the items in it. The midpoints of the classes should not vary widely from the actual averages of the items in the respective classes.

When 160 Illinois farms were grouped by 40-acre classes, with the limits 0–39, 80–119, etc., the items in each class were not equally distributed throughout the intervals (table 5). In the class 80–119, the midpoint was 100, but the average of the farms in the class was only 84, because a high proportion of the farms contained exactly 80 acres.

When these farms were grouped by 40-acre classes, with the limits 1–40, 81–120, etc., the actual average size of the farms in the class was significantly greater than the midpoint, because a large number of farms was exactly the same size as the upper limit (table 5).

When farms were grouped with the limits 20–59, 100–139, etc., the midpoints agreed closely with the average size of the farms in the class (table 4). The prevailing size of farms for each class was the midpoint of the class.

TABLE 5.—EFFECT OF LOCATION OF CLASS LIMITS ON REPRESENTATIVENESS OF FREQUENCY DISTRIBUTIONS

SIZE OF 160 FARMS, BUREAU COUNTY, ILLINOIS, 1930

Typical farms at lower class limits				Typical farms at upper class limits			
Class interval, acres	Frequency	Actual class average	Midpoint of class	Class interval, acres	Frequency	Actual class average	Midpoint of class
0– 39	3	29	20	1– 40	7	35	21
40– 79	16	57	60	41– 80	49	76	61
80–119	48	84	100	81–120	21	107	101
120–159	20	133	140	121–160	43	157	141
160–199	45	163	180	161–200	14	174	181
200–239	6	217	220	201–240	15	236	221
240–279	14	243	260	241–280	5	263	261
280–319	2	280	300	281–320	2	320	301
320–359	2	320	340	321–360	1	360	341
360–399	1	360	380	361–400	1	400	381
400–439	2	402	420	401–440	1	405	421
440–479	0	—	460	441–480	0	—	461
480–519	0	—	500	481–520	0	—	501
520–559	0	—	540	521–560	1	560	541
560–599	1	560	580				
Total	160	—	—	Total	160	—	—

Where feasible, the class limits should be located at multiples of certain commonly used numbers, such as 2, 5, 10, 100, and the like. They should be located so that the mid-values of the classes are also integers familiar to the mind and easy to manipulate.

In a frequency table, there should be no indeterminate classes with only one limit, such as "under 10" or "over 200."

Common methods of designating class limits are as follows:

I	II	III	IV
$ 500–1,000	$ 500 and under $1,000	$ 500– 999	$ 750
1,000–1,500	1,000 and under 1,500	1,000–1,499	1,250
1,500–2,000	1,500 and under 2,000	1,500–1,999	1,750

In each of the above examples, the class limits may be at the same points. The technique of I leaves to one's judgment the classifying of items whose value is 1,000, 1,500, etc. The usual practice with such class limits is to place one-half of the doubtful items in the class above and one-half in the class below. The difficulty of dividing these items equally can easily be overcome by designating the class limits in some other manner. In II, the class limits are definitely stated, and the only objection is the addition of the words "and under," which lengthen the description of the class. They are somewhat difficult to read and interpret, and they take valuable space in printed form.

TABLE 6.—RELATIVE AND CUMULATIVE FREQUENCY DISTRIBUTIONS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Frequency	Relative or percentage distribution	Cumulative distribution* of			
			Numbers		Relatives	
			Upward	Downward	Upward	Downward
−1,500 to −1,001	1	1.1	1	89	1.1	100.0
−1,000 to − 501	2	2.3	3	88	3.4	98.9
− 500 to − 1	11	12.4	14	86	15.8	96.6
0 to 499	14	15.7	28	75	31.5	84.2
500 to 999	17	19.1	45	61	50.6	68.5
1,000 to 1,499	15	16.9	60	44	67.5	49.4
1,500 to 1,999	10	11.2	70	29	78.7	32.5
2,000 to 2,499	6	6.7	76	19	85.4	21.3
2,500 to 2,999	7	7.9	83	13	93.3	14.6
3,000 to 3,499	0	0.0	83	6	93.3	6.7
3,500 to 3,999	1	1.1	84	6	94.4	6.7
4,000 to 4,499	3	3.4	87	5	97.8	5.6
4,500 to 4,999	1	1.1	88	2	98.9	2.2
5,000 to 5,499	1	1.1	89	1	100.0	1.1
Total	89	100.0	—	—	—	—

* Occasionally, the class-interval descriptions for a frequency distribution cumulated downward are "5,000 or more," "4,500 or more," "4,000 or more," and so on. Likewise, cumulated upward, they would read "less than −1,000," "less than −500," and so on.

Method III is usually interpreted as the equivalent of II. The description 500–999 usually means 500 and under 1,000. This fact is sometimes impressed on the reader by writing the description 500–999.9. Method III has the advantages of both definiteness and brevity.

Occasionally, the classes are designated by their midpoints, as in IV. This leaves to the reader the establishment of the upper and lower limits of the class. Method IV has the advantage of brevity, and establishes the midpoint, which may be taken as a single measure likely to represent the items in the class.

RELATIVE FREQUENCY

In a relative or percentage frequency table, the frequency of each class is expressed as a percentage of the total frequency or number of items in the series. In this type of table, the relative frequencies always add to 100 (table 6, left). In the series of 89 incomes for fruit farms, 17 fell in the class $500–999. In the relative frequency table, 17 was equivalent to 19.1 per cent of the total number of farms (17 ÷ 89 = 0.191).

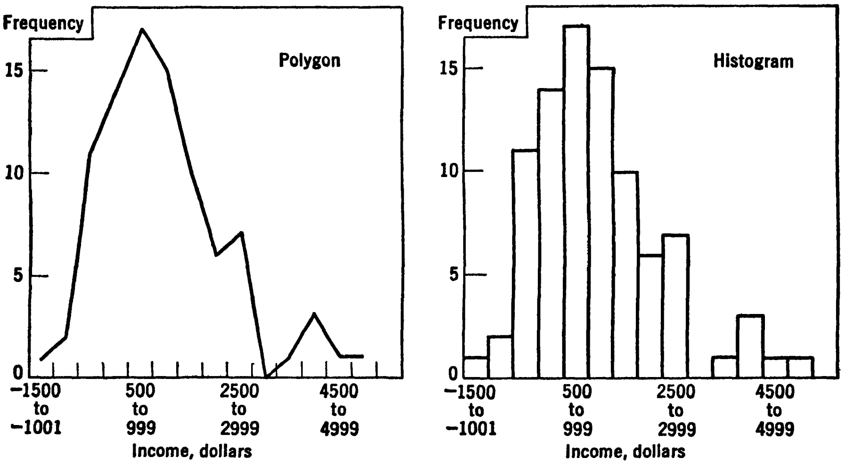


FIGURE 1.—GRAPHIC REPRESENTATION OF A FREQUENCY DISTRIBUTION

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913, CLASS INTERVAL, $500

The polygon (left) is a line connecting the number of farms plotted at the midpoint of the income groups. The histogram (right) is a series of adjacent columns representing the number of farms. The sides of the columns represent the class limits.

CUMULATIVE FREQUENCY

In some frequency tables, it is useful to know not only the frequency of each class, but also the total frequency included in a particular class and in all classes above or below. For example, there were 17 farms in the income class $500–999, and there were 45 farms with incomes of less than $1,000 (table 6, right). There was 1 farm with an income of less than −$1,000 and 3 farms with less than −$500. Frequencies may be cumulative beginning with the smallest values in the series or with the largest. There were 6 farms with incomes of more than $3,000. These cumulative frequencies are sometimes more useful when expressed as percentages. Fourteen farms had minus labor incomes, and these represented 15.8 per cent of the total (table 6).

GRAPHIC REPRESENTATION

The average reader can grasp the characteristics of a frequency distribution presented in graphic form more easily than in tabular form (compare figure 1 and table 2). In a system of rectangular coordinates, the horizontal scale represents the class, and the vertical scale refers to the frequency.

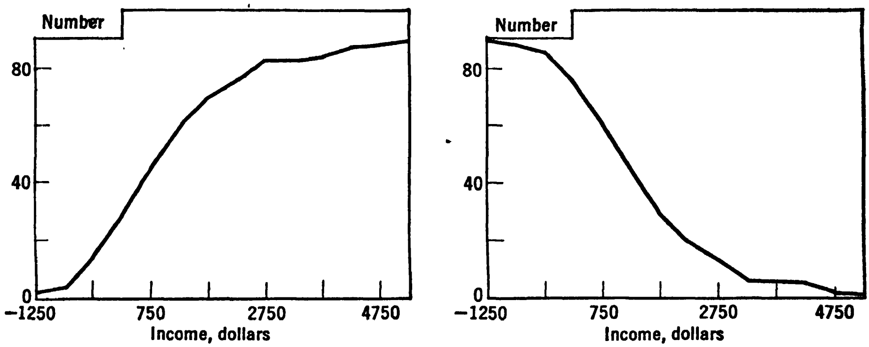


FIGURE 2.—OGIVE OR CUMULATIVE FREQUENCY DISTRIBUTION

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913, CLASS INTERVAL, $500

The curve to the left is a distribution cumulated from the lower incomes; and to the right, from the higher.

The distribution of the labor incomes for 89 fruit farms was represented by (*a*) a frequency polygon and (*b*) a histogram (figure 1). The polygon is the more commonly used method of graphic presentation, but the histogram is probably more correct when the variable is not continuous. The frequency polygon assumes linear interpolation between the midpoints of the classes. The frequencies of the classes are plotted against the midpoints. These plotted points for consecutive classes are connected by straight lines.

The histogram consists of a series of adjoining rectangles whose widths are the class interval and whose lengths are the frequencies. As in the frequency table, each class in a polygon or histogram is designated by its midpoint or by its range.

A cumulative frequency distribution of labor incomes may also be shown graphically (figure 2). This graph which rises from 0 to 89 is called an ogive. It is merely a polygon representing a cumulative rather than a non-cumulative frequency distribution. At any magnitude of the incomes, the ogive indicates the number of incomes less than that amount (figure 2, left), or more than that amount (right). A high frequency is reflected in a steep slope in the curve; a low frequency, in a leveling off of the curve. Ogives are difficult to interpret and are ordinarily of little value to the agricultural economist.

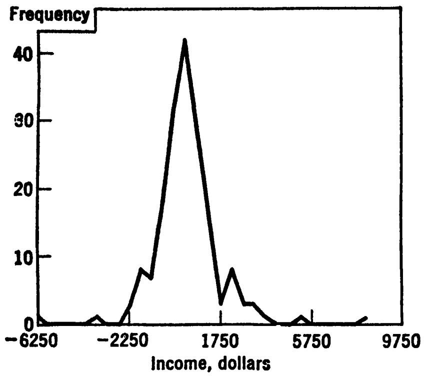


FIGURE 3.—APPROXIMATELY SYMMETRICAL FREQUENCY DISTRIBUTION

Labor Incomes for 178 New York Fruit Farms, 1920, Class Interval, $500

The point of greatest frequency occurs at about the midpoint of the extremes of the data, and there is about the same number of farms above and below the class of greatest frequency.

TYPES OF FREQUENCY DISTRIBUTIONS

Frequency distributions of economic data are of an infinite number of types which may be classified into a few broad categories. The symmetrical distribution is probably the best known but one of the least common types of distribution found in economic phenomena. The normal distribution is a specialized type of symmetrical distribution.

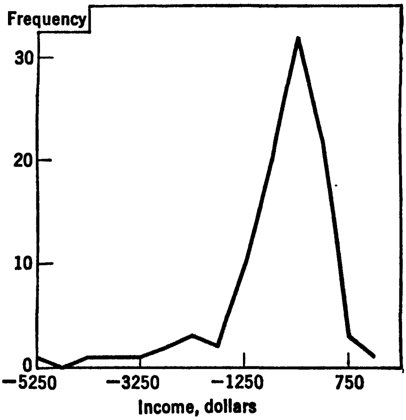


FIGURE 4.—FREQUENCY DISTRIBUTION SKEWED TO THE LEFT

Labor Incomes for 99 New York Fruit Farms, 1914, Class Interval, $500

The point of greatest frequency occurs to the *right* of the midpoint of the range of incomes. There are about 26 farms to the right and 41 to the left of the most frequent class.

The labor incomes of fruit farms for 1920 are a good illustration of an approximately symmetrical distribution (figure 3). In a symmetrical distribution, the average of the series is in the class which has the greatest frequency. The most typical labor income is the average for the series. In this symmetrical distribution, there are approximately equal numbers of farms above and below the class of greatest frequency.

There are also approximately equal numbers of extremely high and extremely low incomes.

Distributions of incomes are often not symmetrical, but skewed or asymmetrical. In a skewed distribution, such as for the 99 fruit farms in 1914, there are unequal numbers of farms on either side of the class of greatest frequency (figure 4). This particular distribution is said to be skewed to the left, or toward the low incomes. In this type of distribution, the class of greatest frequency may be thought of as the most typical income, but, because there are more farms below this class than above it, the average income of the series is below the most typical. In 1914, there were no farms that had incomes greater than $1,500 above the most frequent group, while there were 9 farms that had incomes less than $1,500 below the most frequent class.

The incomes for New York fruit farms in 1918 form an excellent illustration of a frequency distribution that is skewed to the right, or toward the higher incomes (figure 5). In 1918, there were 80 incomes higher than the most usual, and only 31 incomes lower.

Extremely asymmetrical distributions are sometimes termed J-shaped, from the shape of the frequency polygon. In a J-shaped curve, the point of greatest frequency is at or very near one of the extremes of the data. The distribution of the number of hens on 141 farms showed that the most common number was 0 to 199 (figure 6). This indicates that probably the greatest number of the 141 farms kept poultry chiefly for home use and for helping with the grocery bill. As the number of hens increased, the number of flocks decreased. The distribution of size of flocks of strictly commercial poultry farms would probably not have this degree of asymmetry.

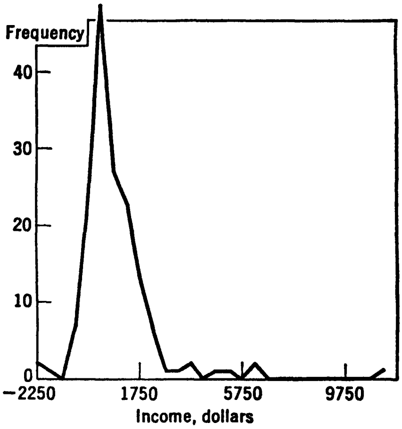


FIGURE 5.—FREQUENCY DISTRIBUTION SKEWED TO THE RIGHT

LABOR INCOMES FOR 159 NEW YORK FRUIT FARMS, 1918, CLASS INTERVAL, $500

The point of greatest frequency occurs to the *left* of the midpoint between the extreme incomes. There are 9 farms that made at least $2,000 more than the most frequent and 2 that made less than $2,000 less.

In a multi-modal distribution, there is no single point of greatest concentration. This type of distribution may be due to an insufficient number of items or too many classes, or to the inherent characteristics of the data. In the former case, the multi-modal condition may be eliminated by increasing the number of items or decreasing the number of classes. In some series, increasing the items or decreasing the classes serves only to accentuate the multimodal condition. When this is true, there is usually some peculiar factor bearing upon the distribution in addition to strictly random fluctuations. In the distribution of acreages of Illinois farms, the frequencies centering around 80- and 160-acre farms are highest, not because of faulty class intervals or insufficient data, but because of the tendency to divide middlewestern lands into multiples of 40 and 80 acres (figure 7). The most usual sizes of farms in the area to which the data refer were 80 and 160 acres. In the group of 50 farms of 60–99 acres, 37 were exactly 80 acres in size. In the group of 50 farms of 140–179 acres, 33 were exactly 160 acres. In the group of 15 farms of 220–259 acres, 11 had exactly 240 acres each.

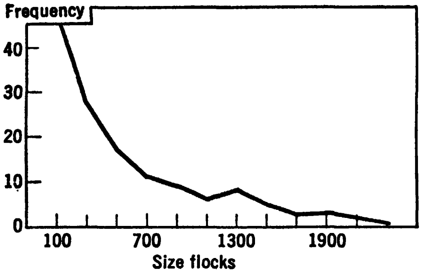


FIGURE 6.—J-SHAPED FREQUENCY DISTRIBUTION

SIZE OF FLOCKS ON 141 NEW YORK FARMS, 1938, CLASS INTERVAL, 200 HENS

As the size of the flocks increased, the number of flocks steadily decreased.

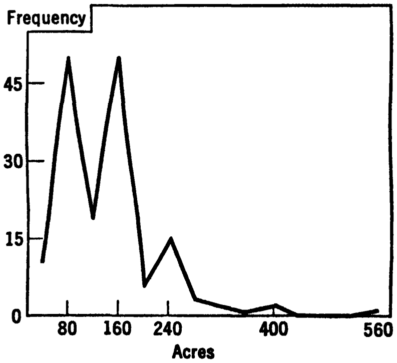


FIGURE 7.—MULTI-MODAL FREQUENCY DISTRIBUTION

SIZE OF 160 FARMS IN BUREAU COUNTY, ILLINOIS, CLASS INTERVAL, 40 ACRES

There are two most frequent groups centering around 80- and 160-acre farms.

Of course, in any multi-modal distribution, two or more points of greatest frequency may be reduced to one if the class interval is increased sufficiently. Where the multi-modal characteristic is inherent in the data, no classification should be used which eliminates it from the frequency polygon. When acreages of 160 farms were classified by 50-acre intervals, 80- and 160-acre farms were included in consecutive classes, and two points of greatest frequency were apparently converted to one (table 4). The 160- and 240-acre farms were not included in consecutive classes, and the frequency of the 170–219-acre class was below that of the 220–269 or of the 120–169-acre classes.

Another general type of frequency distribution has been called U-shaped, again from the shape of the frequency polygon. This type has two points of greatest frequency, each at the extremes of the series. In a sense, it is a specialized type of multi-modal distribution. In another sense, it is a combination of two J-shaped distributions. In a true U-shaped distribution, the tendency to high frequencies at extreme values is inherent in the phenomena, and is not due to random fluctuation. Consequently, a U-shaped distribution cannot be transformed into another type by changing the class intervals or class limits. Various aspects of weather exhibit a tendency to U-shaped distributions. The hours of sunshine at Ithaca, New York, were expressed as a percentage of the total possible for each day. There were numerous days with little or no sunshine, and also numerous days when the sun shone most of the day (figure 8). There were relatively few days when the sun shone as much as 25 to 50 per cent of the possible hours. This phenomenon was due to climatological conditions rather than to chance.

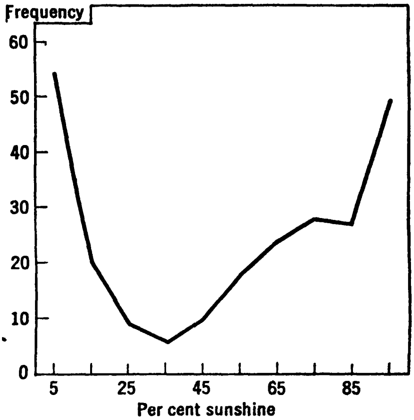


FIGURE 8.—U-SHAPED FREQUENCY DISTRIBUTION

PERCENTAGE OF POSSIBLE HOURS THAT THE SUN SHONE, ITHACA, NEW YORK, MAY TO DECEMBER 1938, CLASS INTERVAL, 10 PER CENT

The days tended to be predominantly clear or predominantly cloudy. There were relatively fewer days when the sun shone approximately one-half of the time.

COMPARISON OF FREQUENCY DISTRIBUTIONS

Two or more frequency distributions may be compared by plotting the frequency polygons on the same chart. Where the numbers of items differ greatly, the distributions may be made more comparable by plotting their percentage frequencies. In this type of comparison, any change due to passage of time or to geographical location, age, sex, type of farm, and the like is usually quite obvious. The most easily identified changes are those in the points of greatest frequency and in the total range of the series.

A frequency distribution of the changes in the top price of cattle from Thursday to Friday shows the most prevalent amount of change between these days and also the range of the changes. It also indicates that the chances of a more-than-usual decline in price are greater than the chances of a less-than-usual decline or of an increase.

A frequency polygon of the changes in top prices from Friday to Monday shows that the most usual changes were small. Further examination indicates that the chances of an increase were greater than the chances of a decrease. The differences between the movements from Thursday to Friday and from Friday to Monday become more striking and obvious to the reader when the two polygons are plotted on the same coordinates (figure 9). Although there was little difference between the most frequent changes in the two instances, there was an important difference in the relative prevalence of advances or declines in prices. If it is assumed that the type of cattle on the market was the same, the chances of an advance in price were greater from Friday to Monday than from Thursday to Friday.

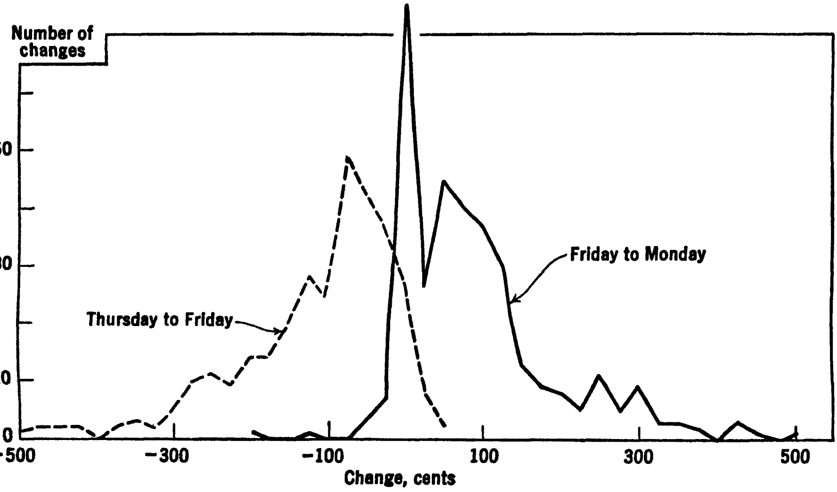


FIGURE 9.—COMPARISON OF TWO DISSIMILAR SKEWED FREQUENCY DISTRIBUTIONS

NUMBER OF CHANGES IN THE PRICE OF TOP CATTLE FROM THURSDAY TO FRIDAY AND FROM FRIDAY TO MONDAY AT CHICAGO, 1924–1929, CLASS INTERVAL, 25 CENTS

The frequencies of price changes were decidedly skewed. The majority of price changes from Thursday to Friday were declines; from Friday to Monday, advances. There was little difference between the most common changes in the two instances.

In like manner, labor incomes in different agricultural areas may be compared. Relative frequency distributions in the Cotton Belt and Corn Belt were found to be quite similar when their polygons were plotted on the same charts (figure 10). The two distributions show one notable difference. Among the Cotton Belt farms, the concentration in the most frequent income group was higher than among the Corn Belt farms. This indicates that there was less variability in incomes of the Cotton Belt than in those of the Corn Belt.

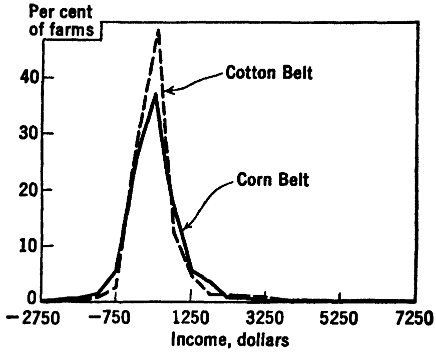


FIGURE 10.—COMPARISON OF TWO SIMILAR FREQUENCY DISTRIBUTIONS

LABOR INCOMES IN THE CORN AND COTTON BELTS, PRE-WORLD WAR I YEARS, CLASS INTERVAL, $500

The two distributions are very similar, and the largest number of farms in both belts made incomes from $1 to $500. The number of farms with great losses or large profits was small in each area, but was largest in the Corn Belt.

USES

Classifying data into frequency distributions is one of the best and easiest ways of arranging a large amount of data in an orderly fashion. The classes of the variable are arranged in order of size. The mass of the original series is condensed into a few classes, the midpoints of which are taken to describe the items. Items which are nearly alike are grouped together in an effort to simplify the series.

Certain characteristics of the series of data may be learned much more readily from a frequency distribution than from the original items.

A frequency table is one of the best condensed summaries that can be made of the data. A large amount of material can be presented in a relatively small amount of space; and, if the table has been properly constructed, it will show most of the characteristics of the series.

Changes due to passage of time, seasonal differences, geographical differences, systems of agriculture, time of the day or week, or to other factors often are vividly shown by plotting two or more frequency polygons on the same coordinates or arranging two or more frequency distributions in the same table. In the absence of mechanical equipment for computations, frequency distributions decrease the amount of "busy" work necessary in calculating various statistical measures.

CHAPTER 2

MEASURES OF CENTRAL TENDENCY

Measures of central tendency are the most common type of statistical measures used to characterize series of data. One speaks of the *average* production per cow, the *average* price of hogs, and the *average* size of farms in Nebraska. There is considerable range in each of the above factors, but the most common characterization is not variability but central tendency.

In the discussion of frequency distributions in chapter 1, both the total range of the data and the points of greatest frequency were pointed out for several distributions. The more conspicuous characteristic in each case was the location of the most frequent class. Even from the array of all the items in the series, the value of the mid-item was almost as obvious as the values of the extremes. The simple or arithmetic average of the series, the value of the point of greatest frequency, and the value of the midpoint of the series are all used as measures of central tendency.

The most common measures of central tendency are: the arithmetic mean, the median, the mode, the geometric mean, and the harmonic mean. Other, less common types are the contra-harmonic mean and the quadratic mean.

ARITHMETIC MEAN

By far the most important type of average is the arithmetic mean, commonly known as the "average." Its calculation is relatively simple from either ungrouped data or frequency tables. To obtain the arithmetic mean of a series of individual items, sum all the items and divide the total by the number of items (table 1). The average labor income of fruit farms was $1,212. This was obtained by adding all incomes and dividing their sum by 89.

FROM FREQUENCY DISTRIBUTIONS

When the number of items in the series is large and mechanical equipment is not readily available, the labor of calculating the arithmetic mean can be considerably reduced through grouping the data in frequency distributions. The midpoint of each class is taken to be the value of each item that falls into that class.

TABLE 1.—CALCULATION OF THE ARITHMETIC MEAN, INDIVIDUAL ITEMS, METHOD I

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Individual items,* X		Calculations
\$ 1,372	.	Arithmetic mean $= \frac{\text{Sum of 89 incomes}}{\text{Number of incomes}}$
− 587	.	
403	.	
1,167	618	$Ma = \frac{\Sigma X}{N} = \frac{\$107,869}{89}$
.	661	
.	735	
Sum	107,869†	$= \$1,212$

* The 89 incomes are given in detail in table 1, page 1.

† In this illustration, there were 14 farms with minus labor incomes aggregating \$4,670 which must be subtracted from the total of 75 farms with plus incomes, \$112,539, to obtain the total for the 89 farms, \$107,869.

Several different methods involve the use of the midpoints of the classes in calculating the arithmetic average. In one method, the midpoint of each class is multiplied by the frequency of the class, and these products are added to find the sum of all the items. This sum is divided by the total number of items to determine the arithmetic mean. The 15 labor incomes in the class \$1,000 to \$1,499 were valued at \$1,250 each (table 2). The midpoint of the class, \$1,250, was multiplied by the frequency, 15. The sum of the products of the midpoints of the classes times their frequencies, \$106,250, divided by 89 gave the arithmetic mean, \$1,194. This value does not check with the previous mean, \$1,212, calculated by Method I, using ungrouped data. This discrepancy may be traced to the inaccuracy introduced in the assumption that the midpoint of each class represented the items in that class, or, specifically, was exactly equal to the arithmetic mean of all the items in that class. In practice, the midpoint of a single class may be quite different from the actual average of the items in the class, but midpoints too high and too low throughout the distribution tend to be compensating; and the error in calculating the arithmetic mean from grouped data is not usually great. This is especially true when there is a large number of items and when the class limits and class intervals have been properly chosen.

TABLE 2.—CALCULATION OF THE ARITHMETIC MEAN, GROUPED DATA, METHOD II

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Midpoint m	Frequency f	Product fm	Calculations*
−1,500 to −1,001	−1,250	1	− 1,250	
−1,000 to − 501	− 750	2	− 1,500	
− 500 to − 1	− 250	11	− 2,750	
0 to 499	250	14	3,500	$Ma = \frac{\Sigma fm}{N}$
500 to 999	750	17	12,750	
1,000 to 1,499	1,250	15	18,750	
1,500 to 1,999	1,750	10	17,500	
2,000 to 2,499	2,250	6	13,500	
2,500 to 2,999	2,750	7	19,250	$= \frac{\$106,250}{89}$
3,000 to 3,499	3,250	0	0	
3,500 to 3,999	3,750	1	3,750	
4,000 to 4,499	4,250	3	12,750	
4,500 to 4,999	4,750	1	4,750	= $1,194
5,000 to 5,499	5,250	1	5,250	
Total	—	89	106,250	

$$* \text{ Arithmetic mean} = \frac{\text{Sum of frequencies of incomes times midpoints of classes}}{\text{Number of incomes}}.$$

The calculation of the arithmetic mean from the frequency table by Method II was much shorter than from individual items by Method I. The transition from Method I to Method II was an important labor-saving step. Additional refinements in method further shortened the calculations. These improvements involve estimating the probable position of the arithmetic mean and calculating a correction which, when added to the "assumed," "estimated," or "guessed" mean, results in the true mean. The first step in these methods is to guess the group in which the arithmetic mean occurs. The arbitrary origin or "assumed mean" is usually the midpoint of the class judged to contain the arithmetic mean. A common procedure is to take the midpoint of the class of greatest frequency as the arbitrary origin. In determining the correction which is added to the arbitrary origin to determine the arithmetic mean, the procedures differ slightly. In Method III, the midpoint used as the arbitrary origin is the basis of comparison. The deviations, D, of each of the other midpoints from the arbitrary origin, A, are calculated by subtracting the arbitrary origin from them.

The next step consists of multiplying the deviations, D, of each class from the arbitrary origin, A, by the frequency, f, of that class. These products of frequencies times deviations, fD, are then summed for the entire series. The correction, c, is determined by dividing the sum of the frequencies times deviations, ΣfD, by the total number of items in the series, N. The arithmetic mean, Ma, is determined by adding the correction, c, to the arbitrary origin, A.

TABLE 3.—CALCULATION OF THE ARITHMETIC MEAN, METHOD III

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Midpoint m	Frequency f	Deviation D	Product fD	Calculations*
−1,500 to −1,001	−1,250	1	−2,000	− 2,000	$Ma = A + c$
−1,000 to − 501	− 750	2	−1,500	− 3,000	
− 500 to − 1	− 250	11	−1,000	−11,000	$c = \frac{\Sigma fD}{N}$
0 to 499	250	14	− 500	− 7,000	
500 to 999	750	17	0	0	
1,000 to 1,499	1,250	15	500	7,500	$A = \$750$
1,500 to 1,999	1,750	10	1,000	10,000	
2,000 to 2,499	2,250	6	1,500	9,000	
2,500 to 2,999	2,750	7	2,000	14,000	$c = \frac{\$39,500}{89}$
3,000 to 3,499	3,250	0	2,500	0	
3,500 to 3,999	3,750	1	3,000	3,000	
4,000 to 4,499	4,250	3	3,500	10,500	$c = \$444$
4,500 to 4,999	4,750	1	4,000	4,000	
5,000 to 5,499	5,250	1	4,500	4,500	$Ma = \$750 + \$444$
Total	—	89	—	39,500	$= \$1,194$

* Arithmetic mean = Arbitrary origin + Correction.

$$\text{Correction} = \frac{\text{Sum of frequencies of incomes times deviations from arbitrary origin}}{\text{Number of incomes}}.$$

In the example of labor incomes on 89 fruit farms, the arbitrary origin, A, was the midpoint of the class \$500–999 (table 3). This midpoint was chosen rather than some other because this class had the greatest frequency and because one might expect from superficial observation that the arithmetic mean would be closer to its midpoint, \$750, than to any other midpoint. The midpoint of the next class above was \$1,250. The deviation of this midpoint from the arbitrary origin was \$500 (1,250 − 750 = + 500). The deviation for the next higher midpoint was \$1,000. The deviation of the next class lower than the arbitrary origin was − \$500; and the next lower, − \$1,000. These deviations from the arbitrary origin in terms of dollars were multiplied by the number of incomes in the respective classes, and these products were summed for the 14 groups of incomes. This sum of frequencies times deviations, + $39,500, was divided by 89, the number of incomes, to determine the correction, + $444. The arithmetic mean was $1,194, the sum of the arbitrary origin, $750, and the correction, $444 (table 3). It may be noted that the arithmetic mean calculated from this frequency distribution was $1,194 regardless of whether Method II or Method III was used.

TABLE 4.—CALCULATION OF THE ARITHMETIC MEAN, METHOD IV

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Midpoint m	Frequency f	Deviation d	Product fd	Calculations*
−1,500 to −1,001	−1,250	1	−4	− 4	$Ma = A + c$
−1,000 to − 501	− 750	2	−3	− 6	
− 500 to − 1	− 250	11	−2	−22	
0 to 499	250	14	−1	−14	$c = \left(\frac{\Sigma fd}{N}\right)i$
500 to 999	750	17	0	0	
1,000 to 1,499	1,250	15	1	15	
1,500 to 1,999	1,750	10	2	20	$c = \frac{79}{89} \times \$500$
2,000 to 2,499	2,250	6	3	18	
2,500 to 2,999	2,750	7	4	28	$= 0.8876 \times \$500$
3,000 to 3,499	3,250	0	5	0	
3,500 to 3,999	3,750	1	6	6	
4,000 to 4,499	4,250	3	7	21	$= \$444$
4,500 to 4,999	4,750	1	8	8	
5,000 to 5,499	5,250	1	9	9	$Ma = \$750 + \$444$
Total	—	89	—	79	$= \$1,194$

$$\text{* Correction} = \left(\frac{\text{Sum of frequencies of incomes times deviations in class intervals}}{\text{Number of incomes}}\right)\left(\text{Class interval}\right).$$

Further refinement in the calculation of the arithmetic mean from a frequency distribution may be obtained by using deviations from an arbitrary origin in terms of the number of class intervals. In this procedure, the arbitrary origin is necessarily the midpoint of some class. The deviation of the class above the arbitrary origin is always + 1, instead of the amount of the class interval. The deviation of the third class below the arbitrary origin is − 3 (table 4). The deviation of each class is multiplied by its frequency, and these products are summed for the whole distribution. This sum is divided by the number of observations and multiplied by the class interval to obtain the correction. As in the preceding method, the correction is added to the arbitrary origin to determine the arithmetic mean.

In the calculation of the arithmetic mean of labor incomes by Method IV, the same arbitrary origin, $750, was used (table 4). The deviation for the $1,250 class was + 1, rather than + $500 as in Method III; the deviation for the next higher class was + 2, and so on. Similarly, the deviations for the lower classes were – 1, – 2, and so on. The frequency for each class was multiplied by the deviation in terms of class intervals for that class, and the sum of these products, 79, was divided by the number of items, 89. The quotient, 0.8876, was multiplied by $500, the class interval. The product is the correction, $444,

TABLE 5.—CALCULATION OF THE ARITHMETIC MEAN WITH DIFFERENT ARBITRARY ORIGINS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Arbitrary origin, $250				Arbitrary origin, $2,250			
Midpoint *m*	Frequency *f*	Deviation *d*	Product *fd*	Midpoint *m*	Frequency *f*	Deviation *d*	Product *fd*
−1,250	1	−3	− 3	−1,250	1	−7	− 7
− 750	2	−2	− 4	− 750	2	−6	− 12
− 250	11	−1	−11	− 250	11	−5	− 55
250	14	0	0	250	14	−4	− 56
750	17	1	17	750	17	−3	− 51
1,250	15	2	30	1,250	15	−2	− 30
1,750	10	3	30	1,750	10	−1	− 10
2,250	6	4	24	2,250	6	0	0
2,750	7	5	35	2,750	7	1	7
3,250	0	6	0	3,250	0	2	0
3,750	1	7	7	3,750	1	3	3
4,250	3	8	24	4,250	3	4	12
4,750	1	9	9	4,750	1	5	5
5,250	1	10	10	5,250	1	6	6
Total	89	—	168	Total	89	—	−188

Calculation:

$$
\begin{aligned}
Ma &= \$250 + \left(\frac{168}{89} \times \$500\right)\\
&= \$250 + (1.8876 \times \$500)\\
&= \$250 + \$943.80\\
&= \$1,194
\end{aligned}
$$

Calculation:

$$
\begin{aligned}
Ma &= \$2,250 - \left(\frac{188}{89} \times \$500\right)\\
&= \$2,250 - (2.1124 \times \$500)\\
&= \$2,250 - \$1,056.20\\
&= \$1,194
\end{aligned}
$$

which was added to the arbitrary origin, $750, to obtain the arithmetic average, $1,194 (table 4).

Identical arithmetic means are obtained from grouped data regardless of which of the above methods is used (tables 2, 3, and 4). Method IV, which involves deviations from the arbitrary origin in terms of class intervals, requires the least effort and is the most common.

SHIFTING THE ARBITRARY ORIGIN

It was pointed out that the midpoint of the labor income class of greatest frequency, $500–999, was used as the arbitrary origin. This practice, though advisable, is not necessary for the accurate calculation of the arithmetic mean. Any midpoint or even any other number may be used as the origin. When the arbitrary origin of the distribution was placed at $250 and at $2,250, the arithmetic means were the same (table 5). They also agreed with the arithmetic mean obtained in table 4.

SIZE OF CLASS INTERVAL

Calculating statistical measures from frequency distributions rather than from ungrouped data introduces some inaccuracy.[1] This inaccuracy varies among different groupings of the same data. The arithmetic mean of labor incomes, $1,194, calculated from a frequency distribution with $500 class intervals did not check with that calculated from distributions of the same data using other class intervals. The mean labor incomes obtained with three different-sized class intervals were:

CLASS INTERVALS	ARITHMETIC MEAN
$ 250	$1,206
500	1,194
2,000	1,225

These may be compared to the actual average, $1,212, from ungrouped data (table 1). The inaccuracy in the use of frequency distributions usually increases with an increase in the size of the class interval and a decrease in the number of classes.

CHARACTERISTICS

The advantages of an arithmetic mean compared with other measures of central tendency are:

1. It is the most easily calculated.
2. It is by far the most commonly used.
3. It is the most easily understood because its calculation is simple and it is the most widely used.

---

[1] Page 17.

4. It is based on all the observations.
5. It is a calculated value, and not based on position in the series.
6. It is adapted to algebraic treatment.

The disadvantages of the arithmetic mean are:

1. Since it includes all the items, its value may be distorted by extreme values.
2. The average is not always a good measure of central tendency, as for instance in extremely asymmetrical distributions.

Other characterisitics are:

1. The sum of the deviations about the arithmetic mean is zero.
2. The sum of the squares of the deviations about a point is a minimum when that point is the arithmetic mean.
3. Its standard error is less than that of any other measure of central tendency.

USES

The arithmetic mean is the most important and the most widely used statistical measure of any kind. The uses of the arithmetic mean are as many and as varied as the activities of man. A discussion of economic questions is full of examples of arithmetic means, or averages, as they are more commonly called. A few of these are: the average yield of wheat in Kansas, the average miles of automobile travel per gallon of gas, the average daily prices of hogs in Chicago, the average cost of producing onions on muck land, the average assessed value of personal property of Illinois farmers, and the average fire loss on Ohio farms. There are myriads of others.

MEDIAN

After the arithmetic mean, the median is one of the most important measures of central tendency. The median is the value of the mid-item of a series arranged in order of size. From ungrouped data, it is a designated value rather than a calculated measure. The procedure in determining the value of the median item is relatively simple. Arrange the items of the series in order of magnitude. When the number of items is odd, the median is the value of the middle item. To determine the middle item, count from one end of the array to the $(N + 1)/2$ item. If the number of observations is even, the median is indeterminate. In this case, the median is arbitrarily the arithmetic average of the values of the two middle observations, the $N/2$ and $(N + 2)/2$ items.

The 89 labor incomes were arranged according to size (table 6). The median was the value of the 45th item, $965. There were 44 incomes less than $965; and 44, greater. *If there had been one more farm with an income of, say, $2,000, the number of items would have been 90,* and the median would have been the average of the 45th and 46th items, or $1,023 [(965 + 1,080) ÷ 2 = 1,023].

TABLE 6.—DETERMINATION OF THE MEDIAN FROM ARRAYED DATA*

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Individual items arrayed from lowest to highest

$ −1,407	$ − 89	$255	$618	$ 866	$1,202	$1,471	$1,965	$2,735
− 587	− 85	259	661	907	1,271	1,523	2,031	2,820
− 573	− 75	290	728	951	1,272	1,543	2,111	2,871
− 467	− 46	387	735	961	1,348	1,585	2,194	2,897
− 328	1	403	735	*965*	1,372	1,748	2,204	3,702
− 316	9	416	801	1,080	1,409	1,867	2,347	4,013
− 206	17	421	819	1,108	1,425	1,879	2,390	4,127
− 194	40	498	845	1,167	1,444	1,887	2,544	4,136
− 186	62	535	854	1,171	1,452	1,900	2,620	4,673
− 111	216	577	855		1,463	1,948	2,732	5,205

* From table 1, page 1.

From grouped data, it is usually impossible to pick out the median item. However, the class containing the median item is easily located. Within the range of this class, the value of the median may be determined by interpolation. The usual method involves a linear interpolation. The observations in the median class may be divided into those above and those below the median item. The proportion of those below the median item to the total frequency of the class is an expression in class intervals of the distance from the lower limit of the class to the position of the median item. The median is the sum of the lower limit of the class plus this proportion of the class interval. The calculation of the median from the lower limit of the class may be shown diagrammatically as follows:

$$\text{Median} = \left[\begin{array}{c}\text{Lower limit}\\ \text{of class}\\ \text{containing}\\ \text{median}\end{array}\right] + \left[\frac{\left(\dfrac{\text{Number of observations}}{2}\right) - \left(\begin{array}{c}\text{Number of items}\\ \text{below the}\\ \text{median class}\end{array}\right)}{\begin{array}{c}\text{Number of items in}\\ \text{the median class}\end{array}}\right]\left[\begin{array}{c}\text{Class}\\ \text{interval}\end{array}\right]$$

The median may also be calculated from the upper limit of the median class.

The median, $985, interpolated from the frequency distribution by the first method was the lower limit of the median class, $500, plus the interpolated amount of the class interval, $485 (table 7).

The median as interpolated from the frequency distribution, **$985**, was not the same as the median from the ungrouped data, **$965** (tables 6 and 7). The interpolated median is only an estimate, and it varies in different frequency distributions of the same data. The error due to interpolation is dependent upon the size of the class interval, the position of the class limits, the number of items in the median class, and the nature of the distribution.

TABLE 7.—APPROXIMATION OF THE MEDIAN FROM A FREQUENCY DISTRIBUTION

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Frequency f	Calculations*
−1,500 to 1,001	1 1	
−1,000 to − 501	2 3	$Me = l_{-i} + \left(\frac{\frac{N}{2} - f_{-i}}{f_0}\right)i$
− 500 to − 1	11 14	
0 to 499	14 28	
500 to 999	17 45	
1,000 to 1,499	15 60	$Me = 500 + \left(\frac{44.5 - 28}{17}\right)(500)$
1,500 to 1,999	10	
2,000 to 2,499	6	
2,500 to 2,999	7	
3,000 to 3,499	0	$= 500 + \left(\frac{16.5}{17}\right)(500)$
3,500 to 3,999	1	
4,000 to 4,499	3	
4,500 to 4,999	1	$= 500 + 485$
5,000 to 5,499	1	
Total	89	= $985

* Me is the symbol for the median.
l_{-i} is the lower limit of the class containing the median.
f_{-i} is the number of items below the median class.
f_0 is the frequency of the median class.
i is the class interval.

The median may be read from an ogive or cumulative frequency polygon such as given in figure 2, page 9. The median may be obtained by drawing a horizontal line from the vertical axis at the median point, $N/2$, and dropping a vertical line to the horizontal axis from the point at which the horizontal line cuts the ogive. At the point on the horizontal scale which is cut by the vertical line, the value of the median may be read. Although this method of determining the median is often pointed out, it is rarely used and deserves little discussion.

CHARACTERISTICS

The advantages of the median compared with other measures of central tendency are:

1. It is easily understood.
2. It is easily determined.
3. Its position is based on all the observations.
4. Its value is not affected by extreme observations at either end of the frequency distribution.
5. The median can be determined regardless of whether the first and last classes in a frequency distribution are indeterminate.

The disadvantages of the median are:

1. It is not a calculated value.
2. It is neither so familiar nor so widely used as the arithmetic mean.
3. The observations of a series must be arranged before the median can be determined.
4. It is not adapted to algebraic treatment.
5. It is erratic if the number of items is small.

Other characteristics are:

1. The numbers of positive and negative deviations from the median are equal.
2. The sum of the first powers of the deviations from the median, without respect to sign, is a minimum.
3. The median has a larger standard error than the arithmetic mean.

USES

The median is a valuable measure of central tendency, but its use is relatively insignificant compared with that of the arithmetic mean. The median is used both as a substitute for and a complementary measure to the arithmetic mean. It is particularly applicable to statistical series with extremely asymmetrical distributions. Since the median is not unduly weighted by the extremely large or small items, it is often a more acceptable measure of central tendency for such series than the arithmetic mean. The median is sometimes used in the preparation of index numbers of prices. At any one time, there may be certain prices which are extremely high or low because of factors other than the price level. If the number of items included in the index is not very large, one or two unusual prices might distort the final index number so that it would not show accurately the movement of prices. The extremely high price of cotton in the northern states during the Civil War unduly affected index numbers of prices when cotton was one of the components.

Measures of central tendency bear certain relationships to each other which change with different degrees of skewness and variability. When the median is used in connection with the arithmetic mean, it affords a method of studying the characteristics of the distribution. A case in point is the comparison of labor incomes on 60 Wisconsin dairy farms in 1916 and 1917. The arithmetic means and medians for the two years were as follows:

YEAR	*Ma*	*Me*
1916	$ 627	$ 515
1917	1,075	1,176

Both measures of central tendency reflected an increase in incomes in 1917 over 1916. For 1916, the arithmetic average exceeded the median. This indicated that there were more farms with less than $627 incomes than with greater incomes. The reason for this was probably a few extremely large incomes; that is, the distribution was skewed toward the larger incomes. In 1917, the situation was reversed. There were more farms with incomes greater than the arithmetic mean, $1,075, than less. The distribution was skewed toward the lower incomes.

MODE

When one speaks of the "man on the street," the "layman," the "typical farm," the "most common wage," and the like, he is unconsciously referring to modes. The mode is the most common item of a series.

The mode differs from the arithmetic mean and the median in that it cannot be determined from a series of ungrouped data.

In a frequency distribution, the mode is the point of greatest concentration; that is, it is the value which is most frequent or most typical for the entire distribution. The class of greatest frequency is usually termed the modal class, and some point within that class is designated as the mode. The exact location of the mode within this modal class is usually determined by some method of approximation. Commonly, the position of the mode is fixed either side of the midpoint of the modal class, depending on the respective frequencies of the classes adjacent to the modal class. If the frequency of the class above is greater than that of the class below, the mode is above the midpoint of the modal class. The procedure is to add to the lower limit of the modal class the proportion of the class interval indicated by the ratio of the frequency of the class above to the sum of the frequencies of the class above and the class below the modal class. Diagrammatically, the mode may be defined as follows:

$$\text{Mode} = \left[\begin{array}{c}\text{Lower limit}\\ \text{of modal}\\ \text{class}\end{array}\right] + \left[\frac{\begin{array}{c}\text{Frequency of class}\\ \text{above modal class}\end{array}}{\left(\begin{array}{c}\text{Frequency of}\\ \text{class above}\\ \text{modal class}\end{array}\right) + \left(\begin{array}{c}\text{Frequency of}\\ \text{class below}\\ \text{modal class}\end{array}\right)}\right]\left[\begin{array}{c}\text{Class}\\ \text{interval}\end{array}\right]$$

The mode, approximated from the frequency distribution of labor incomes by "working up," was the lower limit of the modal class, \$500, plus that proportion of the class interval given by the ratio of 15 to 29 (table 8). The mode was this proportion, \$259, plus \$500, or \$759.

TABLE 8.—APPROXIMATION OF THE MODE

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Midpoint m	Frequency f	Calculations*
−1,500 to −1,001	−1,250	1	$Mo = l_{-i} + \left(\frac{f_{+1}}{f_{-1} + f_{+1}}\right)i$
−1,000 to − 501	− 750	2	
− 500 to − 1	− 250	11	
0 to 499	250	14	
500 to 999	750	17	$Mo = 500 + \left(\frac{15}{14 + 15}\right)(500)$
1,000 to 1,499	1,250	15	
1,500 to 1,999	1,750	10	
2,000 to 2,499	2,250	6	
2,500 to 2,999	2,750	7	$= 500 + \left(\frac{15}{29}\right)(500)$
3,000 to 3,499	3,250	0	
3,500 to 3,999	3,750	1	
4,000 to 4,499	4,250	3	
4,500 to 4,999	4,750	1	$= 500 + 258.62$
5,000 to 5,499	5,250	1	
Total	—	89	$= \$758.62$

* Mo is the symbol for the mode.

f_{-1} and f_{+1} are the number of items in the classes next below and next above the modal class.

i is the class interval.

The value of the mode may be determined from frequency curves. The "true" mode is the value of the variable corresponding to the highest point on the frequency curve which gives the best fit to the actual distribution. In determining the mode, several types of curves have been fitted to the distribution. Some of these are Karl Pearson's mathematical curves, moving averages, and other curves smoothed by some arbitrary method. Unless there is a very large number of items in the distribution, the mode read from a smoothed curve is usually more accurate than the mode approximated from a frequency distribution.

In symmetrical distributions, the arithmetic mean, median, and mode have the same value. It was discovered that, when the distribution was moderately asymmetrical, there was a definite relationship among these three measures, if there were sufficient observations in the distributions. The median lies about one-third of the distance between the arithmetic mean and the mode. From the values of the mean and median in such a distribution, the mode may be estimated by the following formula:

$$Mo = Ma - 3(Ma - Me)$$

This relation can be used to approximate the mode from ungrouped data.

CHARACTERISTICS

The most important advantages of the mode may be summarized as follows:

1. It is by definition the most usual or typical value, and as such is often more representative of the series than any other measure.
2. The mode is not affected by extremely large or small items.

The disadvantages of the mode are:

1. The "true" mode, which is rather rigidly defined, is very difficult to calculate.
2. The "approximate" modes are frequently too inaccurate to be of practical value, especially when a limited amount of data is available.
3. Approximate modes are not adapted to algebraic manipulation.

USES

In the minds of most people, the concept of the mode is the clearest of all the measures of central tendency. When one speaks of an average, he usually means the arithmetic mean; but his conception of the arithmetic mean is frequently that of the mode. The layman assumes no difference between the arithmetic mean and the mode, and often uses "the average" to describe the most usual, most common, or the most typical.

Although the mode may be the most common concept of central tendency, its use in general analysis is almost prohibited by the lack of a satisfactory method of calculation. The arithmetic mean is not hard to calculate and is rigidly defined. Most methods of determining the mode do not define it rigidly, and, as a result, there is too much variability among modes calculated by different methods. Methods of calculation which do define the mode rigidly are prohibitive because of the vast amount of work involved. There is no one method which combines simplicity and accuracy.

In statistical analysis, the mode is sometimes used in preference to other measures of central tendency when the emphasis is placed on the most typical or most common value. The justification for using the mode is that it is the least abstract of all averages and for many purposes is more representative of the data than any other measure.

GEOMETRIC MEAN

The geometric mean differs from the arithmetic mean in that it averages numbers with respect to their geometric rather than arithmetic differences. The arithmetic mean of three numbers is one-third of the sum of the three. The geometric mean of three numbers is the third root, or cube root, of the product of the first number times the second times the third, $Mg = \sqrt[3]{X_1X_2X_3}$. When the number of items is greater than 3 or 4, the tasks of multiplying the numbers and of extracting the root are almost impossible without logarithms. In practice, the geometric mean is found by summing the logarithms of the items and dividing the sum by the number of items. This resulting "average

TABLE 9.—CALCULATION OF THE GEOMETRIC MEAN FROM UNGROUPED DATA

PERCENTAGE OF ALL OWNER-OPERATED FARMS MORTGAGED, 93 COUNTIES OF NEBRASKA, 1930

Original item X	Logarithm $\log X$	Calculations
56.3	1.7505	
62.9	1.7987	$\log Mg = \frac{\Sigma(\log X)}{N}$
82.8	1.9180	
62.0	1.7924	
.	.	
.	.	$= \frac{164.6862}{93}$
.	.	
56.2	1.7497	
51.4	1.7110	
57.8	1.7619	$= 1.7708$
56.9	1.7551	
Total	164.6862	$Mg = 59.0$

logarithm" is the logarithm of the geometric mean, and its value is easily read from a logarithm table.

$$\log Mg = \frac{\log X_1 + \log X_2 + \cdots + \log X_N}{N} = \frac{\Sigma(\log X)}{N}$$

The logarithm of the geometric mean is to the logarithms of the observations as the arithmetic mean is to the observations themselves.

The geometric mean of the percentage of Nebraska farms mortgaged was determined by summing the logarithms of the percentages (table 9). The total of the logarithms, 164.6862, was divided by 93 to obtain the logarithm of the geometric mean, 1.7708. The number corresponding to this logarithm was 59.0, the geometric mean.

The geometric mean may be calculated from a frequency table. As in the calculation of the arithmetic mean, the midpoints are taken to represent the items in the classes. The logarithms of these midpoints are multiplied by their respective frequencies; their products are totaled; and the average is obtained. The natural number corresponding to this average logarithm is the geometric mean, 58.8.

CHARACTERISTICS

The geometric mean has certain advantages among which are:

1. It takes into consideration all the items.
2. It is subject to algebraic manipulation.
3. It is adapted to the manipulation of ratios.

The disadvantages of the geometric mean are:

1. It is not generally understood.
2. It is difficult to calculate.
3. It cannot be determined when there are both negative and positive values in the series, or where one or more of the values is zero.

The geometric mean has additional interesting characteristics:

1. For any series of observations, the geometric mean is always less than the arithmetic mean and greater than the harmonic mean.
2. It gives greater importance to small numbers, and less to large, than does the arithmetic mean. When the distribution is skewed toward the larger values, the geometric mean is often nearer the mode and more typical than the arithmetic mean. Conversely, when the skewness is toward the smaller values, the arithmetic mean is more typical than the geometric mean.

USES

The geometric mean is a relatively unimportant measure of central tendency. The average person does not understand its meaning and has difficulty with its calculation. It is a particularly useful average for variables which tend to increase in a geometric pattern. For example, numbers of some bacteria multiply in a certain ratio to existing numbers of bacteria. The increase in a different period is usually not a constant number but a number based on a constant ratio of the organisms at the beginning of each period.

In the computation of index numbers, the geometric mean has been found to be particularly adaptable to the manipulation of relatives, which are ratios.[2] The geometric mean has a similar application in the comparison of paired variables where the different pairs vary greatly in magnitude. In the following illustration, the taxes paid in 1930 and 1935 were the paired variables. The three farms were not comparable in size, and the arithmetic mean was heavily weighted by farm 1.

FARM	1930	1935
1	$498	$605
2	202	142
3	97	63
Ma	266	270
Mg	214	176

Taxes increased on farm 1 and decreased on farms 2 and 3. The arithmetic mean indicates that the average taxes were practically unchanged, whereas the geometric mean indicates that taxes declined. The arithmetic means reflect the actual dollar changes in the total taxes paid by the three farms. The geometric means reflect the percentage change in the taxes, all farms being weighted equally.

The importance of the geometric mean is much less than is indicated by space devoted to its discussion. The geometric mean is a comparatively unimportant type of average because the human mind is less accustomed to thinking in geometric than in linear differences and is not accustomed to obtaining averages by a process of multiplication. In practice, the geometric mean is rarely used because it is not generally understood, is difficult to determine, and has few specific adaptations.

HARMONIC MEAN

The harmonic mean is a relatively unimportant measure of central tendency, with a restricted application. It is the reciprocal of the arithmetic mean of the reciprocals of the observations. It can be calculated from either ungrouped data or frequency tables. The harmonic mean of the proportion of farms mortgaged was 58.6 from ungrouped data (table 10). The harmonic mean was slightly less than the geometric mean.

---

[2] Page 60.

TABLE 10.—CALCULATION OF THE HARMONIC MEAN FROM UNGROUPED DATA

PERCENTAGE OF ALL OWNER-OPERATED FARMS MORTGAGED, 93 COUNTIES OF NEBRASKA, 1930

Original item X	Reciprocal $1/X$	Calculations*
56.3	0.01776	$\frac{1}{Mh} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_N}}{N}$
62.9	0.01590	
82.8	0.01208	
62.0	0.01613	
.	.	
.	.	$= \frac{\Sigma\left(\frac{1}{X}\right)}{N}$
.	.	
56.2	0.01779	$\frac{1}{Mh} = \frac{1.58697}{93}$
51.4	0.01946	
57.8	0.01730	
56.9	0.01757	$= 0.01706$
Total	1.58697	$Mh = 58.6$

* Mh = harmonic mean.

The formula for the harmonic mean may also be written: $Mh = \frac{N}{\Sigma(1/X)}$.

CHARACTERISTICS AND USES

The harmonic mean has the following advantages:

1. It is based on all observations.
2. It lends itself to algebraic manipulation.
3. It is adaptable to averaging rates of performance.

Some of the disadvantages are:

1. It is not determined by a simple process.
2. It is not easily understood.
3. It greatly magnifies the importance of small numbers.
4. It is meaningless when the observations include both positive and negative values, or when one or more values are zero.

The harmonic mean is primarily used in averaging rates. For instance, if three men take 8, 5, and 4 hours respectively to husk an acre of corn, the problem is to determine the average number of hours to husk an acre. The greater the time required, the slower the speed of husking. The efficiency of the husking is represented by the reciprocals of the time required. The arithmetic mean of 8, 5, and 4 is not the desired answer. The usual procedure is to calculate the acres husked per hour by each man, 0.125, 0.2, and 0.25, find the arithmetic mean of this number of acres, 0.1917, and its reciprocal, 5.22, which is the average hours required to husk an acre. This average is merely the harmonic mean of 8, 5, and 4. The harmonic mean has very little practical application to statistical problems other than those involving rates.

COMPARISON OF "AVERAGES"

Certain definite relationships exist among the various measures of central tendency. Ranked according to size, the means from largest to smallest are: arithmetic, geometric, and harmonic. The geometric and harmonic means are always less than the arithmetic mean, because they give greater weight to the small observations. The sizes of the differences among these three averages depend on the degree of variability in the data.

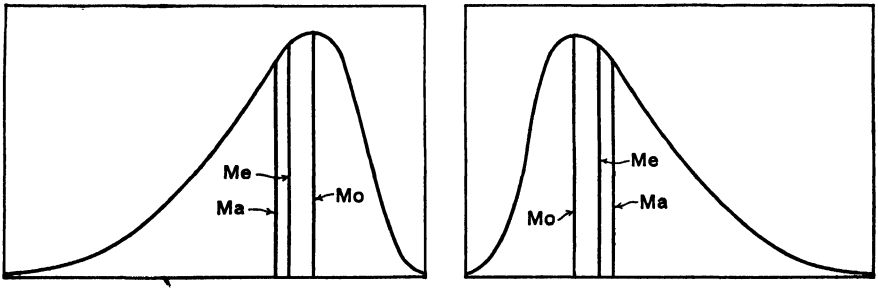


FIGURE 1.—LOCATION OF THE ARITHMETIC MEAN, MEDIAN, AND MODE IN FREQUENCY DISTRIBUTIONS SKEWED TO THE LEFT AND RIGHT

When the direction of skewness is changed, the positions of the arithmetic mean and the mode are reversed. The median is about one-third of the distance from the arithmetic mean to the mode.

The relationship of the median and the mode to each other and to the arithmetic mean depends upon the degree of skewness in the distribution. If it is assumed that there are sufficient observations and that the frequency distribution could be represented by a relatively smooth symmetrical curve, the mode, median, and arithmetic mean have the same values. When the distribution is skewed to the left, the median is less than the mode and the arithmetic mean is less than the median. When skewed to the right, the arithmetic mean is largest, followed by the median and then by the mode (figure 1). The relation of the mode and the median to averages other than the arithmetic mean depends upon their relationship to the arithmetic mean and that of the arithmetic mean to the other measures.

The mode and the median are not calculated from all the observations. The median is based upon position in the series; and the mode, on the degree of concentration. From ungrouped series, the median is a selected value, while all the means are abstractions and often not the same as any one value of the series. The median is the value of the mid-item, and there are as many observations below as above. The mode is the most typical value.

Since the median and the mode are based upon position and relative concentration, respectively, their values are not determined by all the observations. All the calculated averages consider all items in the series.

Of the calculated values, the arithmetic mean is the easiest to determine, and, because it is also easiest to understand, it is by far the most important. All the calculated means are subject to algebraic manipulation, while the approximate mode and median are not.

Only one of the calculated averages, the arithmetic mean, can ordinarily be used when there are both positive and negative values in the series.[3] The mean and mode can be used when there are negative or positive values, or both.

The arithmetic mean is by far the most common and the most useful average. It is well understood, easy to calculate, and is adaptable to all kinds of data. The next most common and most important measures of central tendency are the mode, which is well understood but usually difficult to determine accurately, and the median, which is fairly well understood and relatively easy to determine.

---

[3] All can be used when a measure of central tendency is desired which will ignore sign.

CHAPTER 3

DISPERSION

Averages are the most important measures describing frequency distributions or ungrouped masses of data, but they pertain to only one type of characteristic—central tendency. Other important features are not described by averages. Dispersion, or the nature of the distribution of observations from value to value, must be described by other statistical measures. It is important to know not only the measure of central tendency but also the degree of variability among the items from which it was obtained.

Most of the study of dispersion is confined to variability or the degree of heterogeneity in the values of a series. Minor phases of the study of dispersion are skewness and kurtosis. Skewness deals with the degree of distortion from symmetry present in distributions with a definite degree of concentration. Kurtosis describes the degree of concentration about the mode.

VARIABILITY

Variability is the keynote of world activity. Variability exists in the nature of life, between plants and animals, in orders, families, and species, and even within the species themselves. The method of reproducing life encourages variability and prevents homogeneity. Variability exists in inorganic forms. There are differences due to geography, climatic factors, natural resources, and the passage of the ages. The interaction of all forces produces even greater variability in the pattern of existence. The most important problem of statistical analysis is the study of variability, its amounts and causes. This section is devoted to the measurement of the *amounts* of variability. The range, average deviation, and standard deviation are the more common measures of the amount of variability in data.

Range

The simplest, most easily understood, and most widely used measure of the amount of variability is the range. By the range is usually meant the total range, or the difference between the highest and lowest values in a series. The range can be read very easily from the array in which the values are arranged according to magnitude. The range in labor incomes was $6,612, the difference between −$1,407, the lowest, and + $5,205, the largest (table 1, page 1). The range in data grouped in a frequency distribution may be estimated as the difference between the midpoints of the first and last classes. The range in labor incomes estimated from the frequency distribution was $6,500 (table 2, page 2).

The public is far better acquainted with the range than with any other measure of variability. The use of the range is greatest in the field of price quotations. Since it is physically impossible to print the quotations for all sales for one commodity, let alone for all commodities, a method of abbreviation such as the highest and lowest price is used to indicate the range. Newspapers indicate ranges in the prices of products of interest to their readers. The daily cash price of No. 2 yellow corn at Chicago on October 7, 1940, was quoted as 65 @ 65½¢. The range was ½ cent per bushel.

Ranges in prices may apply to weekly quotations, as well as to daily, like $4 @ 5.25 per 100 lb. of canner and cutter cows at Chicago for the last week of 1938; monthly, like $3.25 @ 5.25 for December 1938; or yearly, like $3.00 @ 6.00 for 1938. Such ranges are most descriptive when they refer to a short period of time, a definite market, and a specific grade of a certain class.

QUARTILE RANGES

There are other types of ranges not commonly known to the layman, but used widely by statisticians. These measures are partial ranges measuring the difference between the values of items at certain positions in the distribution. The median is a measure of central tendency based on position and is the value of the mid-item. The first quartile, Q_1, is the value of the item located one-fourth of the distance from the lowest item to the highest, and the third quartile, Q_3, is located at three-quarters of this distance. The quartiles are directly comparable to the median, which may be called the second quartile. The interquartile range is the difference between the values of the first and third quartiles ($Q_3 - Q_1$). The more common measure of variability is the semi-interquartile range, one-half the distance between the first and third quartiles. This is sometimes known as quartile deviation, $QD = (Q_3 - Q_1)/2$.

The quartiles are calculated in much the same manner as the median. From ungrouped data, the first quartile is the value of the item which divides the series into one-fourth below and three-fourths above. The first quartile is the value of the $\left(\frac{N}{4}+\frac{1}{2}\right)$ item.[1] The first quartile of labor incomes was the value of the 22¾th item $\left(\frac{89}{4}+\frac{1}{2}=22\frac{3}{4}\right)$. Of course, there was no 22¾th item, and so the quartile was interpolated as being three-fourths of the distance between the 22nd and 23rd items (table 1, page 1). Since the values of the 22nd and 23rd items were $259 and $290, respectively, and their difference $31, the value of the 22¾th item would be $282 $(259 + \frac{3}{4} \times 31 = 282)$. The third quartile was the value of the $\left(\frac{3N}{4}+\frac{1}{2}\right)$ item, which was the 67¼th item $\left(\frac{3 \times 89}{4}+\frac{1}{2}=67\frac{1}{4}\right)$. Since the 67th and 68th items were $1,887 and $1,900, respectively, the third quartile was $1,890 $\left(1{,}887+\frac{13}{4}=1{,}890\right)$.

The first quartile, median, and third quartile were the values of the 22¾th, 45th, and 67¼th items. These values were $282, $965, and $1,890, respectively.

The interquartile range was $1,608 (1,890 − 282 = 1,608). The semi-interquartile range or quartile deviation, $804, was one-half this amount. The central half of the incomes was distributed over a range of $1,608, less than one-fourth of the total range. The semi-interquartile range, together with the total range, indicates the amount and nature of the scatter about the median. The size of these two measures indicates the amount of the dispersion, and a comparison of them reveals the nature of the concentration about the median. When the interquartile range is as great as half the total range, there is no tendency in the data to concentrate about a central point. Conversely, the smaller the interquartile range compared to the total range, the greater the concentration about a central point.

Quartile deviation can be calculated from grouped data. The first and third quartiles are interpolated within the classes containing them in the same manner as the median. A diagrammatic representation explains the calculation and indicates the formula.

$$\text{The first quartile} = \left[\begin{array}{c}\text{Lower limit}\\ \text{of class}\\ \text{containing}\\ \text{first quartile}\end{array}\right] + \left[\frac{\left(\frac{\text{Number of observations}}{4}\right) - \left(\begin{array}{c}\text{Number of}\\ \text{items below}\\ \text{quartile class}\end{array}\right)}{\text{Number of items in quartile class}}\right]\left[\begin{array}{c}\text{Class}\\ \text{interval}\end{array}\right]$$

The third quartile may be calculated similarly. The first and third quartiles of labor incomes calculated from grouped data were $295 and

---

[1] The first quartile is sometimes designated as the value of the $\left(\frac{N}{4}\right)$, $\left(\frac{N+1}{4}\right)$, or $\left(\frac{N-1}{4}+1\right)$ item. Many students designate the quartile as the value of the nearest item to the fraction indicated.

TABLE 1.—CALCULATION OF QUARTILE DEVIATION FROM GROUPED DATA

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Frequency f
-1,500 to -1,001	1
-1,000 to - 501	2
- 500 to - 1	11
0 to 499	14
500 to 999	17
1,000 to 1,499	15
1,500 to 1,999	10
2,000 to 2,499	6
2,500 to 2,999	7
3,000 to 3,499	0
3,500 to 3,999	1
4,000 to 4,499	3
4,500 to 4,999	1
5,000 to 5,499	1
Total	89

First quartile,* working up

$$Q_1 = l_{-i} + \left(\frac{\frac{N}{4} - f_{-i}}{f_0}\right)i$$

$$\frac{N}{4} = \frac{89}{4} = 22.25$$

$$Q_1 = 0 + \left(\frac{\frac{89}{4} - 14}{14}\right)(500)$$

$$= 0 + \left(\frac{22.25 - 14}{14}\right)(500)$$

$$= 0 + \left(\frac{8.25}{14}\right)(500)$$

$$Q_1 = 295$$

Third quartile, working up

$$Q_3 = l_{-i} + \left(\frac{\frac{3N}{4} - f_{-i}}{f_0}\right)i$$

$$\frac{3N}{4} = \frac{267}{4} = 66.75$$

$$Q_3 = 1,500 + \left(\frac{\frac{267}{4} - 60}{10}\right)(500)$$

$$= 1,500 + \left(\frac{66.75 - 60}{10}\right)(500)$$

$$= 1,500 + \left(\frac{6.75}{10}\right)(500)$$

$$Q_3 = 1,838$$

Quartile deviation

$$QD = \frac{Q_3 - Q_1}{2}$$

$$= \frac{1,838 - 295}{2}$$

$$= 772$$

Coefficient of quartile deviation

$$V_{QD} = \left(\frac{Q_3 - Q_1}{Q_3 + Q_1}\right)(100)$$

$$= \left(\frac{1,838 - 295}{1,838 + 295}\right)(100) = \frac{154,300}{2,133}$$

$$V_{QD} = 72.34$$

---

* Q_1 and Q_3 are symbols for the first and the third quartiles.
l_{-i} is the lower limit of the class containing the quartile.
f_{-i} and f_0 are the numbers of items below and in the quartile class.
i is the class interval.
QD = quartile deviation or semi-interquartile range.
V_{QD} is the coefficient of variability based on quartile deviation.

\$1,838, respectively (table 1). The quartile deviation was \$772 $\left(\frac{1,838 - 295}{2} = 772\right)$, approximately the same as that from ungrouped data, \$804.

The above measures of variability are in terms of dollars, acres, and the like, and are not comparable from one series of data to another. The relative amount of variability in different series is not shown by their respective quartile deviations. A relative measure for quartile deviation is obtained by dividing the difference between the first and third quartiles by their sum and multiplying by 100. The symbol V_{QD} denotes the coefficient of variability based on quartile deviation. The coefficient of variability[2] in incomes was 72. In the coefficient, the variability is adjusted to the size of the items and thus is comparable from one series to another. This coefficient can never exceed 100 when the values are all positive or all negative numbers. It is of questionable value when the numbers are both positive and negative.

It is possible to calculate other partial ranges, such as the 10–90 percentile range, the 2–8 decile range, and the like. Percentiles are the values of items in the array at those positions which divide the series into 100 equal parts. Similarly, the deciles are the values of items which divide the series into 10 equal parts. The 10–90 percentile range is the difference between the values of the 10th and 90th percentiles. The 10th percentile in labor incomes was −$232. One-tenth of the incomes were less than −$232, and 90 per cent above.

Quartiles, deciles, and percentiles are measures of the values at certain positions and are used in other ways than for the calculation of variability. They are comparable to the median in that they are based on position. They are not measures of central tendency, but are descriptive of the distribution.

These measures of dispersion have been used extensively in the fields of education, psychology, and sociology.

AVERAGE DEVIATION

The more common measures of variability deal with the deviations in the values of the observations from some measure of central tendency —usually the arithmetic mean. The average, or mean, deviation is the arithmetic mean of the deviations of the observations from the arithmetic mean of the series. It is easily calculated from ungrouped data.

Ungrouped Data

By one method, each item is expressed as a deviation from the arithmetic mean, and these deviations are summed without regard to sign. The sum of the deviations divided by the number of items is the average deviation. The average deviation of labor incomes is calculated by subtracting the arithmetic mean, $1,212, from each income, summing these deviations without respect to sign, and dividing this total, $85,711, by 89 (table 2). The average deviation, $963, shows the average amount by which incomes varied from the arithmetic mean, $1,212.

---

[2] The coefficient of variability may be expressed either as a proportion of the whole, 0.72, or as a percentage, 72, with or without the per cent symbol.

TABLE 2.—CALCULATION OF AVERAGE DEVIATION FROM UNGROUPED DATA WITH DEVIATIONS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Labor income X	Deviation from arithmetic mean, $1,212 x	Calculations*
$ 1,372	$ 160	
− 587	−1,799	
403	− 809	$AD = \frac{\Sigma\|x\|}{N}$
1,167	− 45	
.	. .	
.	.	
.	.	$= \frac{85,711}{89}$
− 75	−1,287	
618	− 594	
661	− 551	
735	− 477	AD = $963
Total	|85,711|	

* AD = average deviation.

x = deviations of the original items from the arithmetic mean.

$\Sigma|x|$ = deviations summed without regard to sign.

The average deviation may also be determined from ungrouped data without calculating the individual deviations. The sum of the positive deviations may be found by subtracting from the total of all the values greater than the arithmetic mean the product of the arithmetic mean times the number of items greater. The sum of the negative deviations is the difference between the total of the values lower than the arithmetic mean and the product of the number of items lower, and the arithmetic mean. The sum of these two differences (or, since they are identical, twice one of the differences) divided by the number of items in the series is the average deviation. The method of using only negative deviations may be presented diagrammatically as follows:

$$\text{Average deviation} = \left[\frac{2}{\text{Number of observations}}\right]\left[\left(\begin{matrix}\text{Number of}\\\text{items less}\\\text{than arith-}\\\text{metic mean}\end{matrix}\right)\left(\begin{matrix}\text{Arithmetic}\\\text{mean}\end{matrix}\right) - \left(\begin{matrix}\text{Sum of the}\\\text{items less}\\\text{than the}\\\text{arithmetic}\\\text{mean}\end{matrix}\right)\right]$$

This procedure yields the same average deviation, \$963, as the first method (tables 2 and 3). This method is particularly applicable when the series is very large and tabulating equipment is used.[3]

TABLE 3.—CALCULATION OF AVERAGE DEVIATION FROM UNGROUPED DATA WITHOUT DEVIATIONS

LABOR INCOMES* FOR 89 NEW YORK FRUIT FARMS, 1913

	Incomes less than arithmetic mean, \$1,212 X_{-Ma}	Incomes more than arithmetic mean, \$1,212 X_{+Ma}	Calculations†
	\$-1,407	\$1,271	$AD = \frac{2}{N}[(N_{-Ma})(Ma) - \Sigma X_{-Ma}]$
	- 587	1,272	
	- 573	1,348	
	- 467	1,372	$= \frac{2}{89}(50 \times 1{,}212 - 17{,}745)$
	·	·	
	·	·	
	·	·	$= \frac{2}{89}(60{,}600 - 17{,}745)$
	1,108	4,127	
	1,167	4,136	
	1,171	4,673	$= \frac{2}{89} \times 42{,}855$
	1,202	5,205	
Total	17,745	90,124	$= \frac{85{,}710}{89}$
N	50	39	$AD = \$963$

* From table 1, page 1.

† Ma is the symbol for the arithmetic mean.

N_{-Ma} represents the number of items less than the arithmetic mean.

ΣX_{-Ma} represents the sum of the items less than the arithmetic mean.

Grouped Data

The average deviation may be calculated from a frequency distribution using deviations from the arithmetic mean calculated from that distribution. The deviations of midpoints were multiplied by the frequencies, summed, and averaged. The average deviation was \$966 (table 4). This average deviation from grouped data was practically the same as from ungrouped data, \$963 (table 4 compared with 2 and 3).

The average deviation is usually calculated about the arithmetic mean, but some students have calculated it about other measures of central tendency, notably the median. The average deviation is a minimum when calculated about the median.

---

[3] With such equipment, each term of the diagrammatic formula is easily obtained.

Coefficient of Variability

Average deviations are always expressed in the same unit as the original items. Therefore, the average deviations of two different series, such as incomes and crop yields, are not comparable. Comparable measures of variability can be obtained by expressing the average deviation as a percentage of the arithmetic mean,[4] $V_{AD} = (AD \div Ma)100$. The coefficient of variability for labor incomes was 81 ($966 \div 1{,}194 = 0.809$). This abstract coefficient is comparable with that of any other series.

TABLE 4.—CALCULATION OF AVERAGE DEVIATION FROM GROUPED DATA USING DEVIATIONS FROM THE ARITHMETIC MEAN

LABOR INCOMES* FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Mid-point m	Fre-quency f	Deviations of midpoints from arithmetic mean, \$1,194 $\lvert x\rvert$	Frequency times deviations $f\lvert x\rvert$	Calculations
-1,500 to -1,001	-1,250	1	2,444	2,444	
-1,000 to - 501	- 750	2	1,944	3,888	
- 500 to - 1	- 250	11	1,444	15,884	
0 to 499	250	14	944	13,216	$AD = \frac{\Sigma f\lvert x\rvert}{N}$
500 to 999	750	17	444	7,548	
1,000 to 1,499	1,250	15	56	840	
1,500 to 1,999	1,750	10	556	5,560	
2,000 to 2,499	2,250	6	1,056	6,336	
2,500 to 2,999	2,750	7	1,556	10,892	$= \frac{85{,}944}{89}$
3,000 to 3,499	3,250	0	2,056	0	
3,500 to 3,999	3,750	1	2,556	2,556	
4,000 to 4,499	4,250	3	3,056	9,168	
4,500 to 4,999	4,750	1	3,556	3,556	$AD = \$966$
5,000 to 5,499	5,250	1	4,056	4,056	
Total	—	89	—	85,944	

* Table 2, page 2.

The average deviation is easily understood. It is relatively easily calculated, especially from ungrouped data. The average deviation weights the individual deviations in proportion to their size and does not give undue weight to large or small items, as do some other measures of variability. As a measure of variability alone, the average deviation is unexcelled. Its lack of popularity is due to its unsuitability for algebraic manipulation and for the computation of further statistical measures.

---

[4] V_{AD} represents the coefficient of variability based on the average deviation.

The average deviation is a logical measure of variability, but the standard deviation is more widely used because it can be manipulated algebraically.

STANDARD DEVIATION

The standard deviation is another measure of variability, based on deviations from a central point. The standard deviation is the square root of the average of the squares of the deviations of the items from the arithmetic mean. The difficulty of adding positive and negative deviations without respect to sign is avoided by squaring the deviations.

TABLE 5.—CALCULATION OF STANDARD DEVIATION FROM UNGROUPED DATA BASED ON DEVIATIONS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Labor income X	Deviations from arithmetic mean, \$1,212 x	Deviations squared x^2	Calculations*
\$1,372	\$ 160	\$ 25,600	$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$
−587	−1,799	3,236,401	
403	− 809	654,481	
1,167	− 45	2,025	
.	.	.	$= \sqrt{\frac{137,882,813}{89}}$
.	.	.	
.	.	.	
− 75	−1,287	1,656,369	
618	− 594	352,836	
661	− 551	303,601	$= \sqrt{1,549,245.09}$
735	− 477	227,529	
Total	—	137,882,813	$\sigma =$ \$1,245

* σ, sigma, is almost the universal symbol for the standard deviation.

Ungrouped Data

The standard deviation may be calculated from ungrouped or grouped data. The calculation from ungrouped data follows the same procedure as for the average deviation, except that, in addition, the squares of the individual deviations must be obtained. The calculation is explained in detail by the following diagram:

$$\begin{array}{c}\text{Standard}\\\text{deviation}\end{array} = \sqrt{\frac{\text{Sum of squares of deviations from arithmetic mean}}{\text{Number of observations}}}$$

The standard deviation of labor incomes from ungrouped data was $1,245 (table 5). The standard deviation, like the average deviation, is a measure of central tendency in the variàbility of the observations about the arithmetic mean. The standard deviation is always larger than the average deviation.

From ungrouped data, the calculation is much more difficult for standard deviation than for average deviation, because it involves all the operations necessary for the average deviation and also the squaring of individual deviations. In practice, the standard deviation from ungrouped data is usually not obtained by this method, because the squared deviations are so large that they become unwieldy. The method is included here to aid the beginning student in obtaining a clearer picture of the principles involved in the standard deviation.

When mechanical equipment is available, the standard deviation is often obtained from ungrouped data, but the squares of individual deviations are not calculated. The large amount of work involved in obtaining the individual deviations and their squares is eliminated.

This recently developed method is merely an adaptation of a simple algebraic principle which is important in many phases of modern statistics: The sum of the squares of the deviations about the arithmetic mean is equal to the difference between the sum of the squares of the original items and *N* times the square of the arithmetic mean. This relationship may be shown diagrammatically as follows:

$$\begin{array}{c}\text{Sum of the}\\\text{squares of}\\\text{deviations}\\\text{from}\\\text{arithmetic}\\\text{mean}\end{array} = \left[\begin{array}{c}\text{Sum}\\\text{of the}\\\text{squares}\\\text{of the}\\\text{original}\\\text{items}\end{array}\right] - \left[\begin{array}{c}\text{Number}\\\text{of}\\\text{observa-}\\\text{tions}\end{array}\right]\left[\begin{array}{c}\text{Square}\\\text{of the}\\\text{arithmetic}\\\text{mean}\end{array}\right]$$

and algebraically as follows:

$$\Sigma x^2 = \Sigma X^2 - N(Ma)^2$$

This simple relationship was used to change the method of obtaining the standard deviation as follows:

$$\begin{array}{c}\text{Standard}\\\text{deviation}\end{array} = \sqrt{\frac{\begin{array}{c}\text{Sum of the}\\\text{squares of}\\\text{deviations}\\\text{from arithmetic}\\\text{mean}\end{array}}{\begin{array}{c}\text{Number of}\\\text{observations}\end{array}}} = \sqrt{\left[\frac{\begin{array}{c}\text{Sum of}\\\text{the squares}\\\text{of the}\\\text{original}\\\text{items}\end{array}}{\begin{array}{c}\text{Number of}\\\text{observations}\end{array}}\right] - \left[\frac{\begin{array}{c}\text{Sum of the}\\\text{original}\\\text{items}\end{array}}{\begin{array}{c}\text{Number of}\\\text{observations}\end{array}}\right]^2}$$

or, algebraically,

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

The first diagrammatic and algebraic expressions for the standard deviation are the older methods (table 5); and the second, to the right, the more recent (table 6).

TABLE 6.—CALCULATION OF STANDARD DEVIATION FROM UNGROUPED DATA BASED ON ORIGINAL ITEMS

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Labor income, X	Labor income, squared, X^2	Calculations
$ 1,372	$ 1,882,384	$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$
−587	344,569	
403	162,409	$= \sqrt{\frac{268,621,253}{89} - \left(\frac{107,869}{89}\right)^2}$
.	.	
.	.	
.	.	$= \sqrt{3,018,216 - (1,212)^2}$
618	381,924	
661	436,921	$= \sqrt{3,018,216 - 1,468,944}$
735	540,225	$= \sqrt{1,549,272}$
Total 107,869	268,621,253	$\sigma = \$1,245$

With tabulating equipment, the sum of the original items, ΣX, and the sum of their squares, ΣX^2, are obtained in one series of operations.[5] The standard deviation in incomes calculated by this method was $1,245 (table 6). This method of computation from the squares of the original items is not ordinarily used when tabulating equipment is not available because of the large numbers involved.

Grouped Data

The usual way of avoiding the manipulation of large, unwieldy numbers is by the use of the frequency distribution. Here, again, the method follows that for the average deviation, with the addition of the squares of deviations. The midpoints of the classes may be expressed as deviations in terms of units or class intervals from either the arithmetic mean or an arbitrary origin. When deviations from the arithmetic mean in terms of units are used, the calculation may be explained diagrammatically as follows:

---

[5] Appendix B, page 425.

$$\text{Standard deviation} = \sqrt{\frac{\text{Sum of frequencies times the squares of the deviations of midpoints from arithmetic mean}}{\text{Number of observations}}}$$

The standard deviation in labor incomes calculated by this method was $1,266 (table 7). It is about the same as that from ungrouped data, $1,245.

When the deviations from the arithmetic mean are expressed in units, the squared deviations are likely to be large and unwieldy numbers, as in the example given in table 7.

TABLE 7.—CALCULATION OF STANDARD DEVIATION FROM FREQUENCY DISTRIBUTION AND ARITHMETIC MEAN

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Mid-point m	Fre-quency f	Deviation from arithmetic mean, $1,194 x	Deviations squared x^2	Frequency times deviations squared fx^2	Calculations
−1,500 to −1,001	−1,250	1	−2,444	5,973,136	5,973,136	$\sigma = \sqrt{\frac{\Sigma fx^2}{N}}$
−1,000 to − 501	− 750	2	−1,944	3,779,136	7,558,272	
− 500 to − 1	− 250	11	−1,444	2,085,136	22,936,496	
0 to 499	250	14	− 944	891,136	12,475,904	
500 to 999	750	17	− 444	197,136	3,351,312	
1,000 to 1,499	1,250	15	56	3,136	47,040	
1,500 to 1,999	1,750	10	556	309,136	3,091,360	$= \sqrt{\frac{142,719,104}{89}}$
2,000 to 2,499	2,250	6	1,056	1,115,136	6,690,816	
2,500 to 2,999	2,750	7	1,556	2,421,136	16,947,952	
3,000 to 3,499	3,250	0	2,056	4,227,136	0	
3,500 to 3,999	3,750	1	2,556	6,533,136	6,533,136	
4,000 to 4,499	4,250	3	3,056	9,339,136	28,017,408	$= \sqrt{1,603,585.44}$
4,500 to 4,999	4,750	1	3,556	12,645,136	12,645,136	
5,000 to 5,499	5,250	1	4,056	16,451,136	16,451,136	
Total	—	89	—	—	142,719,104	$\sigma = \$1,266$

Much of the "busy work" and the increasing possibility of mechanical errors may be eliminated by expressing the deviations in terms of class intervals rather than units, and about an arbitrary origin rather than the arithmetic mean.

The sum of the squared deviations about the arithmetic mean is equal to the sum of the squares about any arbitrary origin minus a correction factor which is based upon the difference between the arbitrary origin and the arithmetic mean.[6] This fact is very useful in calculating the standard deviation. When the midpoint of a class is used as an arbitrary origin, the deviations may be computed about the arbitrary origin in terms of class intervals. Since the deviations will be small numbers, like +1, −3, +2, etc., the deviations squared will also be relatively small. Another advantage in using this method lies in the fact that the arithmetic mean is often not known when the student sets out to obtain the standard deviation. Since the correction factor, which is based on the difference between the arbitrary origin and the arithmetic mean, is obtained in the process of calculating the standard deviation, it is not necessary to know the arithmetic mean. The beginning student may not comprehend this method of calculation so easily as the previous methods described, but the results are exactly the same.

---

[6] This correction is the square of that used in the calculation of the arithmetic mean (table 4, page 20).

The various steps in the procedure are as follows:

$$\text{Standard deviation} = \sqrt{\left(\frac{\begin{array}{c}\text{Sum of frequencies}\\ \text{times squares of}\\ \text{deviations from}\\ \text{arbitrary origin in}\\ \text{class intervals}\end{array}}{\text{Number of observations}}\right) - \left(\frac{\begin{array}{c}\text{Sum of the}\\ \text{frequencies}\\ \text{times deviations}\\ \text{in class}\\ \text{intervals}\end{array}}{\text{Number of observations}}\right)^2}\left(\begin{array}{c}\text{Class}\\ \text{interval}\end{array}\right)$$

The standard deviation from grouped data was approximately the same, $1,267 and $1,266, whether calculated by the above method or the preceding one (tables 7 and 8).

TABLE 8.—CALCULATION OF STANDARD DEVIATION FROM FREQUENCY DISTRIBUTION AND ARBITRARY ORIGIN

LABOR INCOMES FOR 89 NEW YORK FRUIT FARMS, 1913

Class interval, dollars	Midpoint m	Frequency f	Deviations in class intervals d	d^2	fd	fd^2	Calculations
−1,500 to −1,001	−1,250	1	−4	16	−4	16	$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}(i)$
−1,000 to − 501	− 750	2	−3	9	−6	18	
− 500 to − 1	− 250	11	−2	4	−22	44	
0 to 499	250	14	−1	1	−14	14	
500 to 999	750	17	0	0	0	0	$= \sqrt{\frac{641}{89} - \left(\frac{79}{89}\right)^2}(500)$
1,000 to 1,499	1,250	15	1	1	15	15	
1,500 to 1,999	1,750	10	2	4	20	40	
2,000 to 2,499	2,250	6	3	9	18	54	$= \sqrt{7.20224719 - (0.8876)^2}(500)$
2,500 to 2,999	2,750	7	4	16	28	112	
3,000 to 3,499	3,250	0	5	25	0	0	
3,500 to 3,999	3,750	1	6	36	6	36	$= \sqrt{7.20224719 - 0.78783376}(500)$
4,000 to 4,499	4,250	3	7	49	21	147	
4,500 to 4,999	4,750	1	8	64	8	64	
5,000 to 5,499	5,250	1	9	81	9	81	$= 2.533 \times 500$
Total	—	89	—	—	79	641	= $1,267

The difference in the size of numbers in the examples of the two methods is impressive (tables 7 and 8). The sums of the columns fx^2 and fd^2 were 142,719,104 and 641, respectively.

As with the calculation of the arithmetic mean, the value of the standard deviation is the same regardless of which arbitrary origin is used. When the arbitrary origin is zero, the deviations in the observations are the observations themselves, and the one step of calculating deviations is eliminated. The determination of the standard deviation in this special case is as follows:

$$\begin{matrix}\text{Standard}\\ \text{deviation}\end{matrix} = \sqrt{\left(\frac{\begin{matrix}\text{Sum of frequencies}\\ \text{times squares of}\\ \text{midpoints}\end{matrix}}{\begin{matrix}\text{Number of obser-}\\ \text{vations}\end{matrix}}\right) - \left(\frac{\begin{matrix}\text{Sum of frequencies}\\ \text{times midpoints}\end{matrix}}{\begin{matrix}\text{Number of obser-}\\ \text{vations}\end{matrix}}\right)^2}$$

Numerous variations of this method would be possible, such as expressing the value of midpoints in terms of class intervals and expressing the midpoints as deviations, not from zero, but from the midpoint of the lowest class.

Many variations in all the methods of calculating standard deviations from frequency distributions may be found in textbooks. Even others are possible. Most of these involve the same basic principles and yield the same result.

Coefficient of Variability

The average labor income was $1,194; and the standard deviation, $1,266 (table 7). The standard deviation is 106 per cent of the average income. This relationship has been termed the coefficient of variability based on standard deviation, and is denoted $V_\sigma = (\sigma \div Ma)100$. The advantages and limitations of this coefficient are the same as those given on page 40 for the coefficient based on quartile deviation.

In studying only labor incomes, the standard deviation, $1,266, is the more valuable measure of variability; but in comparing the variability in incomes with that in size of farms, yields per acre, number of cows, and the like, the coefficient of variability, 106, is the more valuable.

Comparison of Measures of Variability

Range and partial ranges, such as quartile deviation, are "position" measures whose values depend upon those of items in definite positions in the array. They are not affected by all the observations, and indicate nothing regarding the distributions between their limits. Average and standard deviations are "calculated" values, based upon all the observations.

Standard deviation is always greater than the average deviation, because the squaring of the deviations before averaging gives the larger deviations greater weight. Both the standard and average deviations are greater than quartile deviations. The average deviation is usually about four-fifths of the standard deviation, while the quartile deviation is about two-thirds of the standard deviation. A range of one standard deviation on either side of the arithmetic mean usually includes about two-thirds of the observations. As indicated by the definition, the range of one quartile deviation on either side of the mean includes about one-half of the observations. The average deviation lies between the standard and quartile deviations, and its corresponding range includes between one-half and two-thirds of the observations.

TABLE 9.—RELATIONSHIPS AMONG THREE MEASURES OF VARIABILITY FOR A NORMAL DISTRIBUTION

Measures of variability	Percentage of observations included within a given range on either side of the mean			Size of each measure of variability relative to standard deviation
	± one deviation	± two deviations	± three deviations	
Quartile deviation.......	50.0	82.3*	95.7*	0.6745
Average deviation.......	57.5	88.9*	98.3*	0.7979
Standard deviation......	68.3	95.4	99.7	1.0000

* Not commonly used.

For a normal distribution, these relationships are fixed and have been determined mathematically (table 9). The exactness of these relationships is disturbed when the distribution departs from symmetry or normality. In distributions that are not normal, the *order of size* of the three measures is unchanged, but the *relative size* is changed. In symmetrical distributions, which are not normal, their relative sizes change with the type of concentration about the average.

The average deviation has a great advantage over other measures of variability in that it is easily understood. Students recognize it as a simple arithmetic average. Ranges and partial ranges, like the quartile deviation, are fairly well understood. The standard deviation is difficult for beginning students to comprehend. They often calculate it without understanding its underlying principle.

In the ease of calculation, quartile deviation ranks ahead of average deviation, and the standard deviation is a poor third, when the data are ungrouped. There is little difference in the time required for the determination of the three measures from frequency distributions.

The standard deviation is the only measure which is well adapted to algebraic treatment. Quartile deviation is not a calculated measure, and the average deviation is awkward to express algebraically because plus and minus signs are considered alike in its calculation.

In an elementary treatment of statistical series in which a measure of variability is desired only for itself, any one of the three would be acceptable. Probably the average deviation would be superior. However, in usual practice, the measure of variability is employed in further statistical analysis. For such a purpose, the standard deviation is by far the preferable. The standard deviation lends itself to the analysis of variability in terms of the normal curve of error. Practically all advanced statistical method deals with variability and centers around the standard deviation. This alone explains why the standard deviation has been used to a far greater extent than the other measures of variability.

USES

Elementary statistical analysis often ends with the calculation of the average or the standard deviation. Most students are interested in the direct application of such measures to the problem at hand. The value of the most important statistical measure, the arithmetic mean, lies in its use for making comparisons. Measures of variability are also of direct value for comparisons. They are useful in comparing the amount of dispersion in several series. Sometimes, the coefficient of variability is more useful than the original measure, especially when the different series are not in the same units.

The yield of corn in various states and for different years has often been studied by the use of averages. A more revealing analysis could be made with measures of variability. Averages tell nothing about the way the yield in Iowa varies from year to year, or the size of year-to-year fluctuations in Iowa compared with those in Nebraska. For the 50-year period, 1880–1929, the average deviation of corn yield for Iowa was 4.6 bushels per acre; and for Nebraska, 5.7 bushels (table 10). These average deviations show variability in terms of bushels. Because the average yield was not the same in all states, average deviations cannot be compared directly. When these average deviations were expressed in percentage of their respective means, the resulting coefficients of variability were comparable. The coefficients of variability were 12 for Iowa and 21 for Nebraska. Variability in the yield of corn is less in the

TABLE 10.—VARIABILITY IN THE YIELD OF CORN PER ACRE IN SIX CORN BELT STATES, 1880–1929

State	Average yield *Ma*	Average deviation* *AD*	Coefficient of variability V_{AD}
Iowa	37.2	4.6	12
Illinois	34.5	4.5	13
Indiana	34.1	4.3	13
Missouri	28.1	3.8	14
Nebraska	27.7	5.7	21
Kansas	21.9	6.3	29

* The variability could have been measured by the standard deviation in place of the average deviation. The differences among states would have been about the same.

"heart" of the Corn Belt than in areas farther south and west. The most important cause of this is the difference in climate.

The variation in corn yields may be compared with the variation in yield of other crops in the same states. Because wheat yields are lower than yields of corn, the measure used for comparisons among crops for the six states was the coefficient of variability (table 11). For all three crops, corn, oats, and wheat, variability in yield was least in the northern states, Wisconsin and Minnesota, a little higher in Iowa and Illinois, and greatest in Kansas and Nebraska (table 11). Again, the cause for these differences was climate. The high variability in Nebraska and Kansas was due to the greater prevalence of drought during the growing season.

TABLE 11.—COEFFICIENT OF VARIABILITY IN THE YIELDS* OF CORN, OATS, AND WHEAT IN SIX MIDWESTERN STATES, 1880–1929

State	Corn	Oats	Wheat
Wisconsin	11	10	14
Minnesota	12	14	15
Iowa	12	13	17
Illinois	13	14	17
Nebraska	21	16	20
Kansas	29	21	19

* Based on average deviation.

Measures of variability may also be useful in farm-management studies. A group of 680 farms in Illinois were classified as to tenure;

and the size of farms, receipts, expenses, incomes, and profits were studied. Comparisons between owner-operated and share-rented farms indicated more variability in the former (table 12). The difference in variability between the two types of farms was small for size, somewhat greater for receipts, expenses, and farm incomes, and greatest for net profits.

Averages for these factors indicated that the share-rented farms were larger in size and had the greater farm receipts and expenses and farm incomes. In spite of their smaller size, owner-operated farms were more variable in size, in method of operation, and in profits. Regardless of type of tenure, farm receipts, expenses, and income were more variable than size of farm; labor income was still more variable; and profits were extremely variable.

Prices are a fertile field for the study of variability. Variability in prices of individual commodities at a given time or variability in the price of one commodity with the passage of time may both be studied advantageously. Variability in the price structure with rising, falling, and stable prices has held the interest and imagination of many students.

TABLE 12.—COEFFICIENTS OF VARIABILITY* FOR SIZE AND INCOMES OF FARMS OPERATED BY OWNERS AND SHARE TENANTS

680 Farms in McHenry County, Illinois, 1912

Factor	Owners	Share tenants
Size of farm........	44	34
Farm receipts.......	59	38
Farm expenses......	63	42
Farm income.......	78	60
Labor income.......	169	115
Net profit..........	381	132

* Based on standard deviation.

In the field of marketing, one might study variability in the cost of different methods of marketing, variability in sales from time to time, from store to store, city to city, and the like. The home economist has studied variability in consumption of various foods and articles of clothing according to age, racial groups, income levels, and the like.

Space prevents listing more of the many fields in which the amount of variation may be measured, comparisons made, and valuable conclusions reached.

A great deal of statistics is concerned with the relationships among different series. Methods of analyzing these relationships usually involve a measure of variability. This measure has practically always been the standard deviation or its square. The greatest use of the standard deviation has been in the analysis of relationships, rather than in the study of a single variable.

SKEWNESS

Skewness is another characteristic of the frequency distributic Standard and average deviations are measures of the amount of d persion. Skewness describes the nature of this dispersion. Symmetri distributions are not skewed. In asymmetrical or skewed distributio the number of observations either side of the mode is unbalanced. distribution is said to be skewed to the right when more than one-h the observations are greater than the mode. Conversely, it is skewed the left when more than one-half are smaller than the mode.

A wide variety of measures of skewness have been developed.[7] Th are abstract coefficients generally based on relative positions of quartile, median, mode, and arithmetic mean. They differ greatly the ease of calculation. Their values are not comparable.

Most statistical workers in the field of economics have not b concerned with the problem of skewness.[8] They have been more in ested in the amount of variability than in its nature. The difficulty calculating the coefficient of skewness usually outweighs its usefuln

KURTOSIS

Kurtosis is another characteristic of the frequency distribution. L skewness, it also describes the nature of dispersion. Kurtosis is degree of peakedness in the distribution or in the concentration at central tendency. A distribution more peaked than the normal cu is said to be *leptokurtic;* less peaked, *platykurtic;* and normal, *mesoku* Kurtosis[9] is an unimportant statistical measure, and for all pract purposes may be ignored.

---

[7] Karl Pearson developed a simple coefficient of skewness based on the diffe between the arithmetic mean and the mode in terms of standard deviations:

$$Sk = \frac{Ma - Mo}{\sigma}$$

[8] Mills, F. C., The Behavior of Prices, Publications of the National Bure Economic Research, No. 11, p. 574, 1927, used coefficients of skewness in a of the dispersion among prices of individual commodities during periods of and falling price levels.

[9] Kurtosis = $\beta_2 - 3$, where $\beta_2 = \frac{\Sigma x^4}{N} \div \left(\frac{\Sigma x^2}{N}\right)^2$.

CHAPTER 4

INDEX NUMBERS

An index number[1] is a comparative measure of magnitude. It is a ratio of the magnitude of a variable at one time, place, or position to its magnitude at another. It may also be the ratio of one variable to another. Index numbers indicate changes and differences, and they are very useful tools in the study of prices and other variables.

The simplest index number involves the ratio of only two things. For example, the ratio of the United States farm price of corn in 1932 to the price in 1926 was 0.402, or an index of 40.2

$$\left(\frac{\text{1932 price corn}}{\text{1926 price corn}} = \frac{28.1¢}{69.9¢} = 0.402, \text{ or index } 40.2\right)$$

Index numbers are commonly expressed as percentages. The 1932 price of corn was 0.402 times, or 40.2 per cent of, its 1926 price.

In terms of wheat, the index of the 1932 price of corn was 72.4 ($28.1 \div 38.8 = 0.724$, a ratio, or 72.4, an index). A common example of an index is the so-called corn-hog ratio—more accurately the hog-corn ratio—which is the bushels of corn required to equal in value 100 pounds of hogs. This ratio is not expressed as a percentage, but as a proportion. In 1932, the corn-hog ratio was 12.3 ($3.47 \div 0.281 = 12.3$).

The ratio of the price of hogs in Georgia to the price in Iowa is an index number based on geographical differences. The index for 1933 was 111:

$$\frac{\text{Georgia price}}{\text{Iowa price}} = \frac{\$3.00}{\$2.70} = 1.11, \text{ or index} = 111$$

A great variety of these simple index numbers, or relatives as they are sometimes termed, is in constant use. It is a common trait of the human mind to think of magnitude in terms of relative rather than absolute values. The size of a given variable at a given time or place bears meaning only when it is compared to the size of another variable or to the same variable at a different time or place. To state that the price of corn in 1932 was 28.1 cents might mean little to one who did not know the prices previous to or following 1932. To state that the price in 1932 was about 40 per cent of the price in 1926 or 43 per cent of the 1910–1914 average price is to indicate that the price in 1932 was much lower than in either of the other two periods.

---

[1] The words "indices," "indexes," and "index numbers" are commonly used for the plural of "index." These expressions were derived from the Latin word *index* related to the Latin *indico*, meaning "point out." Since the expressions "index number" and "index numbers" are rather long, the words "index," "indexes," and "indices" are more frequently used. The word "indices" is the Latin plural for "index"; "indexes" is the Anglicized plural.

The most common type of index number measures prices with passing time. The base period for index numbers is the period with which others are compared, and is usually fixed. Index numbers of prices are usually percentages of prices for the base period.

Index numbers of individual commodities are relatively simple, easily understood, and readily calculated, but of little interest to the student of statistics. The prices of various products for a given period may be combined into a single index number. Group index numbers are less important than individual indexes, but more difficult to calculate, much more difficult to understand, and, consequently, more interesting to the statistician.

A person can easily grasp the movement in prices of a single commodity. However, the mind has great difficulty in averaging the movements of a large number of prices, some of which may be rising and others falling. The combination of a number of single indexes into one combined index greatly simplifies comparisons.

The index for 1932 farm prices in terms of 1926 may be calculated for a part of or for all farm products. Many different methods of combining these prices have been devised. Some of these are relatively simple; others are difficult. The advantages of the difficult methods are not usually sufficient to compensate for their disadvantages. A few relatively simple methods are adequate for most statisticians.

UNWEIGHTED INDEX NUMBERS

Sum of Numbers, or Simple Aggregative

One of the simplest methods of combining prices and calculating index numbers involves the simple addition of the price quotations. The sum of the price quotations for any period expressed as a percentage of the sum of the corresponding quotations in the base period is the index number. This method is called a sum of numbers, or a simple aggregative. The prices of a bushel of corn and of wheat, a pound of butter and of cotton, and 100 pounds of hogs totaled $9.158 in 1910–1914, $14.417 in 1926, and $4.408 in 1932 (table 1). One procedure is to consider each of these totals as index numbers indicating the relative level of prices for the three periods. Another practice is to express each total as a percentage of the totals for the base period. When this base period was specified as 1910–1914, the index for 1926 was 157.4; and for 1932, 48.1 (4.408 ÷ 9.158 × 100 = 48.1). When 1926 was the base period, indexes were 63.5 for 1910–1914, 100 for 1926, and 30.6 for 1932.

TABLE 1.—INDEX NUMBERS BASED ON SUMS OF NUMBERS

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	1910–1914	1926	1932
Corn, per bu.	$0.648	$ 0.699	$0.281
Wheat, per bu.	0.880	1.351	0.388
Butter, per lb.	0.256	0.416	0.211
Hogs, per 100 lb.	7.250	11.800	3.470
Cotton, per lb.	0.124	0.151	0.058
Total	$9.158	$14.417	$4.408
Index, 1910–1914 = 100	100	157.4	48.1
Index, 1926 = 100	63.5	100	30.6

The sum-of-numbers method is one of the easiest to understand and to use, but it contains a serious fault. Since the size of the physical unit for each product affects its quotation, the importance of products quoted by large physical units is overemphasized. In the above example, the prices total $4.408 in 1932, over three-fourths of which was contributed by the hog quotation, $3.47. Practically nothing was contributed by the cotton, $0.058 (table 1). The same was true for 1926 and for 1910–1914. Changes in the price of hogs affected the index about four times as much as changes in all the other four commodities. The indexes actually showed the changes in hog prices slightly modified by the changes in the other four prices.

The sum-of-numbers method is satisfactory when the physical units are chosen so that their quotations are very nearly the same. When the prices were expressed in terms of 1.5 bushels of corn, 1 bushel of wheat, 3 pounds of butter, 12 pounds of hogs, and 7 pounds of cotton, each product contributed about equally to the index (table 2).

In practice, a group of price quotations is almost never sufficiently uniform to justify the use of the sum-of-numbers method. Although changing the physical units and adjusting their quotations render the method satisfactory, this process destroys one of the great advantages of the sum of numbers, ease of calculation.

TABLE 2.—INDEX NUMBERS BASED ON SUMS OF NUMBERS WITH MODIFIED QUOTATIONS

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Quantity	1910–1914	1926	1932
Corn	1.5 bu.	$0.972	$1.049	$0.422
Wheat	1 bu.	0.880	1.351	0.388
Butter	3 lb.	0.768	1.248	0.633
Hogs	12 lb.	0.870	1.416	0.416
Cotton	7 lb.	0.868	1.057	0.406
Total	—	$4.358	$6.121	$2.265
Index, 1910–1914 = 100		100	140.5	52.0
Index, 1926 = 100		71.2	100	37.0

ARITHMETIC MEAN OF RELATIVES

A common method of calculating the index of a group of prices is to average the index numbers or relatives for the individual commodities. The arithmetic mean is the most common average employed. The index of the price of an individual commodity is the ratio of the price to the corresponding price for the base period. The price relative for corn in 1932 on the 1910–1914 base was 43.4 (0.281 ÷ 0.648 × 100 = 43.4) (table 3). The corresponding relative for hogs was 47.9. The arithmetic mean of the five relatives for 1932 was 52.9, indicating that these prices

TABLE 3.—INDEX NUMBERS BASED ON THE ARITHMETIC MEAN OF RELATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Prices			Relatives		
	1910–1914	1926	1932	1910–1914	1926	1932
Corn, per bu.	$0.648	$ 0.699	$0.281	100	107.9	43.4
Wheat, per bu.	0.880	1.351	0.388	100	153.5	44.1
Butter, per lb.	0.256	0.416	0.211	100	162.5	82.4
Hogs, per 100 lb.	7.250	11.800	3.470	100	162.8	47.9
Cotton, per lb.	0.124	0.151	0.058	100	121.8	46.8
Total	—	—	—	500	708.5	264.6
Arithmetic mean of relatives or indexes, 1910–1914 = 100				100	141.7	52.9

were about one-half of the 1910–1914 average. The arithmetic mean of the 1926 relatives was 141.7.

This method tends to give equal importance to all the products in the index if the price movements are approximately the same. When such is not the case, the relatively highest-priced article contributes more than its share to the index. In 1932, the relative for butter was 82.4, about twice the other relatives. Butter contributed one-third to the 1932 index, while each of the other commodities contributed about one-sixth. In 1926, the relatives for corn and cotton were considerably lower than those for wheat, butter, and hogs. Consequently, corn and cotton influenced the combined index less than the other commodities. Such fluctuations are year-to-year phenomena and would be involved in the construction of any type of index number. They are not faults of the particular method of constructing the index number, but the reasons for their construction. The arithmetic mean of the relatives method does give the various commodities approximately equal weights over a long period of time.[2] It is not difficult to calculate and has been used very widely.

MEDIAN OF RELATIVES

In the discussion of central tendency, it was stated that for distributions containing extremely large or small items the arithmetic mean might not be so typical an average as the median.[3] When prices are changing rapidly, a few prices may precede or lag behind the majority of commodities and may unduly affect the arithmetic mean. There are also times when unusual conditions cause one or two individual prices to rise much higher than the rest. The price of cotton in northern states during the Civil War and the price of potash during World War I are cases in point. Because such situations do exist occasionally, some statisticians prefer to calculate group index numbers on the basis of the median rather than the arithmetic mean.

The 1932 relatives arranged according to size indicate that the median was 46.8 (43.4, 44.1, *46.8*, 47.9, 82.4) (table 3). The median, 46.8, was somewhat lower than the arithmetic mean, 52.9, which was unduly affected by the very high relative for butter, 82.4. The 1926 median index, 153.5, obtained in the same way, was larger than the arithmetic mean of relatives, 141.7.

The median may be somewhat erratic when based on such a small number of relatives. In practice, a much larger group of commodities is usually included in an index.

---

[2] This is true if there is no persistent long-time trend in any of the prices relative to the trend of the group.

[3] Page 26.

GEOMETRIC MEAN OF RELATIVES

The geometric mean of relatives is sometimes designated as the index number of a group. Whereas the arithmetic mean weights equal arithmetic differences alike, the geometric mean weights equal ratio differences alike.[4] The geometric mean emphasizes the small items and discounts the importance of the large ones. Consequently, it is always less than the arithmetic mean, but not greatly different when there is little variability in the relatives.

TABLE 4.—INDEX NUMBERS BASED ON THE GEOMETRIC MEAN OF RELATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	1910–1914		1926		1932	
	Relatives	Logarithms of relatives	Relatives	Logarithms of relatives	Relatives	Logarithms of relatives
Corn	100	2.0	107.9	2.03302	43.4	1.63749
Wheat	100	2.0	153.5	2.18611	44.1	1.64444
Butter	100	2.0	162.5	2.21085	82.4	1.91593
Hogs	100	2.0	162.8	2.21165	47.9	1.68034
Cotton	100	2.0	121.8	2.08565	46.8	1.67025
Total	—	10.0	—	10.72728	—	8.54845
Average log	—	2.0	—	2.14546	—	1.70969
Index, 1910–1914 = 100		100.0	—	139.8	—	51.2

The geometric mean is a root extracted from the product of a group of numbers, and is most easily obtained by the use of logarithms. The geometric mean of relatives is the index whose logarithm is the arith-

---

[4] The relative prices of two products at two periods are:

PRODUCTS	PERIOD I	PERIOD II
A	100	200
B	100	50
Arithmetic mean		125
Geometric mean		100

The arithmetic differences from the arithmetic mean are the same, 75 (200 − 125 = 75, and 125 − 50 = 75). The geometric or ratio differences from the geometric mean are the same, 2 (200 ÷ 100 = 2, and 100 ÷ 50 = 2).

metic mean of the logarithms of the individual items. By this method, the index of prices of five farm products was 51.2 for 1932, and 139.8 for 1926 (table 4). These index numbers were slightly lower than those based on the arithmetic mean because the importance of larger relatives was discounted. They were quite different from the indexes based on the median because the geometric mean is based on all the relatives, while the median is not.

TABLE 5.—INDEX NUMBERS BASED ON THE GEOMETRIC MEAN OF PRICES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	1910–1914		1926		1932	
	Price	Logarithm	Price	Logarithm	Price	Logarithm
Corn	64.8¢	1.81158	69.9¢	1.84448	28.1¢	1.44871
Wheat	88.0	1.94448	135.1	2.13066	38.8	1.58883
Butter	25.6	1.40824	41.6	1.61909	21.1	1.32428
Hogs	725.0	2.86034	1,180.0	3.07188	347.0	2.54033
Cotton	12.4	1.09342	15.1	1.17898	5.8	0.76343
Total	—	9.11806	—	9.84509	—	7.66558
Average	—	1.82361	—	1.96902	—	1.53312
Minus log for 1910–1914		1.82361	—	1.82361	—	1.82361
Log of ratio to 1910–1914		0	—	0.14541	—	9.70951–10
Log index,* 1910–1914 = 100		2.0	—	2.14541	—	1.70951
Index, 1910–1914 = 100		100	—	139.8	—	51.2

* The addition of 2 to the logarithm of the ratio is the same as multiplying by 100, or moving the decimal point two places.

Geometric means of relatives may be calculated more simply (table 5). The step involving the determination of the relatives may be omitted. The ratio of the geometric mean of the quotations to the geometric mean of the corresponding quotations for the base period is identical to the geometric mean of relatives.[5]

---

[5] The geometric mean of relatives of the commodities may be expressed as follows:

$$\sqrt[5]{\frac{\text{Price corn '26}}{\text{Price corn '10–'14}} \times \frac{\text{Price hogs '26}}{\text{Price hogs '10–'14}} \times \text{etc.}}$$

$$\frac{\sqrt[5]{\text{Price corn '26} \times \text{Price hogs '26} \times \text{etc.}}}{\sqrt[5]{\text{Price corn '10–'14} \times \text{Price hogs '10–'14} \times \text{etc.}}}$$

The two expressions are algebraically identical. The statistical procedure based on the first formula is given in table 4; on the second, in table 5.

The calculation of the geometric mean of the quotations involves the tabulation, addition, and averaging of the logarithms of the actual prices for each period in question (table 5). The index numbers, which are the ratios of the geometric averages for the given periods to the geometric average for the base period, may be found by subtracting the average logarithm of the base period from the other average logarithms and finding the number whose logarithm is this difference. The average logarithms for 1932 and 1910–1914 prices were 1.53312 and 1.82361, respectively (table 5). The natural numbers corresponding to these logarithms are geometric means of prices. Since the geometric means are in terms of logarithms, their ratio may be found most easily by subtracting the logarithm for 1910–1914 from that for 1932 (1.53312 − 1.82361 = 9.70951–10). The addition of 2 to the logarithm places the index in percentage terms (9.70951 − 10 + 2 = 1.70951 = logarithm of 51.2). The geometric means of relatives and of prices yield exactly the same index numbers (tables 4 and 5). The method in table 5 takes less time and consequently is the more widely used.

WEIGHTED INDEX NUMBERS

The preceding methods assumed that equal weighting of commodities in final index numbers was desired. Because commodities are not always of equal importance, methods have been devised to give each commodity a specific bearing on the final index proportionate to its importance.

TABLE 6.—DETERMINATION OF WEIGHTS FOR INDEX NUMBERS

Commodity	Amount, 000,000	1910–1914 price	Value, 000	Percentage weights
Corn	513 bu.	64.8¢	$ 332,424	10
Wheat	658 bu.	88.0¢	579,040	18*
Butter	2,004 lb.	25.6¢	513,024	15*
Hogs	122 (100 lb.)	$ 7.25	884,500	27
Cotton	7,970 lb.	12.4¢	988,280	30
Total	—	—	3,297,268	100

* The percentages are 17.561 and 15.559. In order to have the weights total 100, only the former was raised. If the total were to be raised, a similar principle would apply, and a percentage such as 14.4845 might be raised to 15.

WEIGHTED ARITHMETIC MEAN OF RELATIVES

The weights used for the arithmetic mean of relatives may be derived from physical quantities and prices in a variety of ways. The common basis of weights is the value of products in the base or some other specified period. Sometimes the weights are expressed in terms of their original values, and sometimes on a percentage basis.

Five agricultural commodities were weighted according to the value of sales (table 6). The value of corn sold, $332 million, was 10 per cent of the total value of the five commodities and was given a percentage weight of 10. When percentage weights are used, the weighted arithmetic mean of relatives is determined by multiplying each relative times its weight, summing the products, and dividing by 100. The 1932 relative for wheat, 44.1, was multiplied by its weight, 18. The relative for butter, 82.4, was multiplied by its weight, 15 (table 7). The products for wheat, 794, butter, 1,236, and for corn, hogs, and cotton were summed. The total, 5,161, was divided by 100 to obtain the weighted index, 51.6.

TABLE 7.—INDEX NUMBERS BASED ON THE WEIGHTED ARITHMETIC MEAN OF RELATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Percentage weights*	1926		1932	
		Relatives†	Products	Relatives†	Products
Corn	10	107.9	1,079	43.4	434
Wheat	18	153.5	2,763	44.1	794
Butter	15	162.5	2,438	82.4	1,236
Hogs	27	162.8	4,396	47.9	1,293
Cotton	30	121.8	3,654	46.8	1,404
Total	100	—	14,330	—	5,161
Index, 1910–1914 = 100		—	143.3	—	51.6

* Table 6. † Table 3.

The 1932 weighted index, 51.6, was somewhat lower than the unweighted arithmetic mean, 52.9; for 1926 the reverse was true.[6] These small differences were due to the relative weights given commodities which were high or low in the particular years.[7]

For those students who expect to calculate a long series of index numbers, the following procedure would effect a considerable saving of time over the above method.

---

[6] Table 3, page 58, compared with table 7.

[7] In usual practice, as in the above example, weights are expressed as percentages. The values upon which the percentages are calculated could themselves be employed as weights. The sum of the products of these values times the relatives divided by the total value would give the same index numbers. This procedure involves more calculation and consequently is less popular.

The weighted arithmetic mean of relatives may be written diagrammatically as follows:

$$\text{Weighted index} = \sum\left(\frac{\text{Price for given period}}{\text{Base price}} \times \begin{matrix}\text{Respective per-}\\\text{centage weight}\end{matrix}\right)$$

$$= \sum\left(\begin{matrix}\text{Price for given}\\\text{period}\end{matrix} \times \frac{\text{Weight}}{\text{Base price}}\right)$$

With constant weights and a fixed base period, the expression Weight ÷ Base price is a constant for each commodity for the entire period. Therefore, the index is the sum of the products of prices times constant multipliers.

TABLE 8.—THE USE OF MULTIPLIERS IN THE CALCULATION OF THE WEIGHTED ARITHMETIC MEAN OF RELATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Percentage weights	1910–1914	Multipliers	1926		1932	
				Price	Product	Price	Product
Corn..........	10	64.8¢	0.1543	69.9¢	10.8	28.1¢	4.3
Wheat........	18	88.0¢	0.2045	135.1¢	27.6	38.8¢	7.9
Butter........	15	25.6¢	0.5859	41.6¢	24.4	21.1¢	12.4
Hogs..........	27	\$ 7.25	3.7241	\$11.80	43.9	\$ 3.47	12.9
Cotton........	30	12.4¢	2.4194	15.1¢	36.5	5.8¢	14.0
Index, 1910–1914 = 100.....		—	—	—	143.2	—	51.5

The multiplier[8] for corn was the percentage weight, 10, divided by the 1910–1914 price, 64.8, or 0.1543. The multipliers for the other farm commodities are given in table 8. The 1926 and 1932 prices of corn, 69.9¢ and 28.1¢, times the multiplier, 0.1543, give the products 10.8 and 4.3. These, added to the products for the other commodities, give directly the indexes 143.2 for 1926 and 51.5 for 1932 (table 8). These indexes are the same as those given in table 7 except for the difference due to insufficient decimal places in some calculations.

This procedure has the advantage over the alternative method of saving a great deal of time and energy in the calculation of a long series of indexes. The calculation of relatives for each period is eliminated by the calculation of one set of multipliers. This procedure has the disadvantages that it is probably more difficult for the beginning student to follow, and is not a saving of labor when index numbers for only one or two periods are desired.

---

[8] If the prices were quoted in dollars, the multiplier would be 15.43.

WEIGHTED GEOMETRIC MEAN

The same system of percentage weights may be employed in calculating weighted geometric means of relatives or prices. The logarithms of relatives or prices, whichever the case may be, are multiplied by the weights. The sum of the products of weights times the logarithms of relatives divided by 100 gives the logarithm of the index.

TABLE 9.—INDEX NUMBERS BASED ON THE WEIGHTED GEOMETRIC MEAN OF RELATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Percentage weights	1910–1914		1926		1932	
		Log of relatives*	Product, log × weight	Log of relatives*	Product, log × weight	Log of relatives*	Product, log × weight
Corn......	10	2.0	20	2.03302	20.33020	1.63749	16.37490
Wheat.....	18	2.0	36	2.18611	39.34998	1.64444	29.59992
Butter....	15	2.0	30	2.21085	33.16275	1.91593	28.73895
Hogs......	27	2.0	54	2.21165	59.71455	1.68034	45.36918
Cotton....	30	2.0	60	2.08565	62.56950	1.67025	50.10750
Total.....	100	—	200	—	215.12698	—	170.19045
Average...	—	—	2.0	—	2.1512698	—	1.7019045
Index, 1910–1914 = 100.....			100.0	—	141.7	—	50.3

* Table 4.

The simplest procedure is to arrange the weights, relatives, and logarithms of relatives in an orderly manner (table 9). The logarithms of relatives are multiplied by the weights, summed, and averaged. By use of a table of logarithms, the index corresponding to this average logarithm may be readily found, 50.3 for 1932.

When prices rather than relatives are used, the logarithm of the index is found by averaging the products of weights and the logarithms of prices (table 10). Again, the difference between this procedure and that shown in table 9 is the elimination of the price relatives.

TABLE 10.—INDEX NUMBERS BASED ON THE WEIGHTED GEOMETRIC MEAN OF PRICES

Indexes of Prices of Five Farm Products

Commodity	Percentage weights	1910–1914		1926		1932	
		Log of price*	Product, log × weight	Log of price*	Product, log × weight	Log of price*	Product, log × weight
Corn	10	1.81158	18.11580	1.84448	18.44480	1.44871	14.48710
Wheat	18	1.94448	35.00064	2.13066	38.35188	1.58883	28.59894
Butter	15	1.40824	21.12360	1.61909	24.28635	1.32428	19.86420
Hogs	27	2.86034	77.22918	3.07188	82.94076	2.54033	68.58891
Cotton	30	1.09342	32.80260	1.17898	35.36940	0.76343	22.90290
Total	100	—	184.27182	—	199.39319	—	154.44205
Average		—	1.8427182	—	1.9939319	—	1.5444205
Minus log for 1910–1914		—	1.8427182	—	1.8427182	—	1.8427182
Log of ratio to 1910–1914		—	0	—	0.1512137	—	9.7017023–10
Log index 1910–1914 = 100		—	2.	—	2.1512137	—	1.7017023
Index, 1910–1914 = 100		—	100.0	—	141.7	—	50.3

* Table 5.

Weighted Aggregative

For reasons to be discussed later, the so-called weighted aggregative index has gained increasing popularity. This method is merely an extension of the sum of numbers or simple aggregative method involving the application of weights. For each period including the base, the total value of given amounts of commodities is computed. The ratio of this total value for a given period to the total value in the base period is the weighted index number. The weights are physical quantities. They are not based upon values and are not percentages. The weight for corn, 513,000,000 bushels, is multiplied by the 1932 price, 28.1 cents, to obtain the value of corn for that year, $144,000,000 (table 11). The sum of the 1932 values of 513,000,000 bushels of corn, 658,000,000 bushels of wheat, and so on was $1,707 million. Since the same quantities were worth $3,297 million at 1910–1914 prices, the index for 1932 was 51.8 (1,707 ÷ 3,297 × 100 = 51.8). When the same physical weights are used with the weighted arithmetic mean of relatives and with the weighted aggregatives, the two indexes are identical.[9] In such cases, the weighted aggregative method is merely a short process for obtaining the arithmetic mean of relatives. The physical weights which are multiplied by prices in table 11 are proportional to the multipliers used in table 8.

---

[9] Differences are due to insufficient decimals in calculation.

TABLE 11.—INDEX NUMBERS BASED ON WEIGHTED AGGREGATIVES

INDEXES OF PRICES OF FIVE FARM PRODUCTS

Commodity	Physical weight, 000,000 omitted	1910–1914		1926		1932	
		Price	Value, 000,000 omitted	Price	Value, 000,000 omitted	Price	Value, 000,000 omitted
Corn.......	513 bu.	64.8¢	$ 332	69.9¢	$ 359	28.1¢	$ 144
Wheat.....	658 bu.	88.0¢	579	135.1¢	889	38.8¢	255
Butter.....	2,004 lb.	25.6¢	513	41.6¢	834	21.1¢	423
Hogs.......	122 (100 lb.)	$7.25	885	$11.80	1,440	$3.47	423
Cotton.....	7,970 lb.	12.4¢	988	15.1¢	1,203	5.8¢	462
Total......	—	—	3,297	—	4,725	—	1,707
Index, 1910–1914 = 100.....		—	100	—	143.3	—	51.8

COMPARISON OF INDEX NUMBERS

There was considerable variation in the index numbers for 1926 and for 1932 due to methods of calculation. Two general types of unweighted index numbers were prepared: (1) the sum of numbers or simple aggregative, and (2) averages of relatives.

TABLE 12.—COMPARISON OF INDEX NUMBERS OBTAINED BY VARIOUS METHODS

INDEXES OF PRICES OF FIVE FARM PRODUCTS, 1910–1914 = 100

Method	Index numbers		
	1910–1914	1926	1932
Unweighted			
Sum of numbers, simple aggregative (table 1).......	100	157.4	48.1
Sum of numbers, simple aggregative modified (table 2)	100	140.5	52.0
Median of relatives (page 59)....................	100	153.5	46.8
Mean of relatives			
Arithmetic (table 3)..........................	100	141.7	52.9
Geometric (table 4)..........................	100	139.8	51.2
Harmonic (calculations not given)..............	100	137.8	50.0
Weighted			
Mean of relatives			
Arithmetic (table 7)..........................	100	143.3	51.6
Geometric (table 9)..........................	100	141.7	50.3
Harmonic (calculations not given)..............	100	140.0	49.4
Aggregative (table 11).........................	100	143.3	51.8

The sum-of-numbers method usually results in erratic index numbers because the size of physical units is not comparable for all commodities. One hundred pounds of hogs were worth many times more than a bushel of corn or a pound of cotton, and this index was predominantly a reflection of hog prices. Since hog prices increased from 1910–1914 to 1926 relatively more than the other products, the 1926 index by this method, 157.4, was higher than that by any other method (table 12). This fault was remedied when the physical units were adjusted so that quotations were about the same for all commodities. When this procedure was followed, the resulting index number, 140.5, was much less than that given above, 157.4, and practically the same as the arithmetic mean of relatives, 141.7. The advantages of the sum-of-numbers method are its simplicity and ease of calculation. However, when physical units and quotations were modified, these advantages were lost.

The median was also erratic when the number of commodities was small. The 1926 index, 153.5, was higher than all others except the sum of numbers; and the 1932 index, 46.8, was the lowest of all. However, with a large number of commodities, the median would usually be less erratic than other index numbers. In the opinion of many, an index number should show the most typical change in a group. The median satisfies this requirement better than most calculated averages. Extreme variations from the most typical unduly affect the size of all "calculated" index numbers. They affect the median of relatives only as any large item affects the position of the median.[10] After the individual relatives have been obtained and arrayed according to size, the median is easily determined by inspection. The disadvantage of the median is its unsuitability for small groups.

Index numbers which are means of relatives always rank in size from largest to smallest as follows: arithmetic, geometric, and harmonic. The amount of variation among these means depends upon the amount and nature of variability in the relatives.

The arithmetic mean is by far the most important of all unweighted index numbers. It is well understood and is adaptable for groups of any size. It is relatively easy to calculate. The determination of the individual relatives requires considerable time, but this is more of an advantage than a disadvantage because many persons are interested in the relative change of individual commodities as well as of groups.

The geometric mean of relatives, though never of great practical importance, has been the center of much theoretical discussion and controversy. It is more difficult to calculate and much more difficult to understand than the arithmetic mean, but it has the advantage of weighting equal ratio differences alike. The arithmetic mean always weights equal arithmetic differences alike, whether the relative be high or low. When there are a few extremely high relatives, the geometric mean is usually nearer the median and gives a more reliable index than the arithmetic mean. Conversely, when there are a few extremely small relatives, the arithmetic mean may be preferable.

---

[10] For these reasons, the mode of relatives might be a better index than even the median if a simple and accurate method of determining the mode could be devised.

As to method of calculation, the weighted aggregative index is somewhat comparable to the sum of numbers or simple aggregative. However, the resulting weighted index does not have the disadvantage of the sum of numbers. The predominating influence of large quotations in the sum-of-numbers index is usually compensated for in the weighted aggregative by differences in physical units. The modified sum of numbers is really a weighted aggregative. The weighted aggregative is very similar to the arithmetic mean of relatives both in ease of calculation and in the size of the index. The weighted aggregative and weighted arithmetic mean of relatives are identical when the same physical quantities and base period prices are used as a basis of weights. On this basis, the respective 1926 indexes are the same, 143.3; and for 1932, almost the same,[11] 51.8 and 51.6 (table 12).

The relationships among various weighted means of relatives are the same as those among unweighted means of relatives. The weighted arithmetic mean is larger than the geometric or harmonic means; and the harmonic is the smallest. The arithmetic mean is by far the most important weighted-mean index. There is little to choose between the weighted arithmetic mean and the aggregative from the standpoint of calculation, simplicity, comprehensibility, or size of the index. However, the arithmetic mean has been used much more than the weighted aggregative, partly because of custom and partly because of interest in the individual relatives.

The primary advantage sought in the weighting of index numbers is greater accuracy. The ideal of accuracy is never attained in any index number, weighted or unweighted. The quotations themselves are estimates representing an infinitesimal part of the sales of any one or any group of commodities. An index number represents only a sample and is subject to sampling errors. Further inaccuracies arise because of the difficulty of obtaining accurate weights.

---

[11] Differences are due to insufficient decimals in calculations.

For 1926, the difference between the unweighted arithmetic and geometric means was 1.9 points; that between the weighted and unweighted arithmetic means, 1.6 points. In the light of all the variables affecting index numbers, those differences due to weighting were probably negligible. In the attempt to attain perfection, the problem of weighting has been and will probably continue to be given attention out of proportion to its importance. There are, no doubt, cases where the relative importance of the various commodities is so different that weighting would increase accuracy.

The five most common indexes are the weighted and unweighted arithmetic and geometric means of relatives, and the weighted aggregative. The greatest difference between any two of these indexes was 3.5 in 1926 and 2.6 in 1932. The above differences[12] are insignificant compared with the striking change in prices from 1926 to 1932.

The final choice from the above five methods is usually made on the basis of personal preference. In making the choice, the doubtful advantages of greater accuracy of some index numbers must be weighed against the greater ease with which other index numbers are calculated and understood.

TIME-REVERSAL TEST

Several years ago, accuracy in converting a series of index numbers from one base period to another caught the imagination of many students interested in index-number theory. This common practice of conversion was challenged. The usual criterion for testing the validity of this conversion was comparison with the index recalculated on the new base. To convert the 1932 index of five commodities from the 1910–1914 to the 1926 base, the usual practice was to divide the 1932 index by the 1926 index. The converted weighted aggregative index was 36.1 ($51.8 \div 143.3 \times 100 = 36.1$). When the aggregative was recalculated with 1926 as 100, the index was the same, 36.1 (table 13). The weighted aggregative was, therefore, said to satisfy the time-reversal test. Certain common methods, such as weighted and unweighted means and medians of relatives, were severely criticized because they did not meet the test.[13] Much of the little popularity the weighted and unweighted geometric means have enjoyed was due to their convertibility. The sum-of-numbers method also satisfies the test.

For the six methods, the greatest differences between converted and recalculated indexes, 0.2 point, was probably overshadowed by other types of errors (table 13). The practical significance of the convertibility test has been greatly overemphasized.

---

[12] Because of the small number of commodities, five, these differences were greater than otherwise would have been expected.

[13] Weighted or unweighted harmonic means do not satisfy the test.

TABLE 13.—COMPARISON OF CONVERTED AND CALCULATED INDEXES FOR 1932

1926 = 100

Method	Converting the 1932 index from the 1910–1914 to the 1926 base			Calculated 1932 index,† 1926 = 100
	Calculated indexes, 1910–1914 = 100		Converted 1932 index,* 1926 = 100	
	1926	1932		
Sum of numbers	157.4	48.1	30.6	30.6
Arithmetic mean of relatives	141.7	52.9	37.3	37.5
Geometric mean	139.8	51.2	36.6	36.6
Weighted arithmetic mean‡	143.3	51.6	36.0	36.2
Weighted geometric mean‡	141.7	50.3	35.5	35.5
Aggregative‡	143.3	51.8	36.1	36.1

* The converted 1932 index, on the 1926 base, was obtained by dividing the 1932 index, on the 1910–1914 base, by the 1926 index on that base. (For the sum-of-numbers method, 48.1 ÷ 157.4 = 0.306.)

† These indexes were independently calculated from the prices in table 1, page 57.

‡ Based on fixed weights of relatives.

WEIGHTS

In the determination of weights for index numbers, there is the problem of assigning to each component a weight proportionate to its importance in the index. The basis of estimates of importance varies with the subject of the index number.

For example, an index of retail food prices in Chicago should be based upon the amount of the different types of foods purchased by a normal family and not upon the amounts of food produced in the United States, the amounts produced in the area about Chicago, or the amounts coming into or processed in Chicago.

In general, it is better to weight farm prices on the basis of sales rather than production because prices are obtained for sales and not for production, and because there would be no duplication of products. For instance, in an index of Iowa farm prices, if the weights were based on total production, the index would include all the hogs and all the corn produced. Since hogs are merely corn on the hoof, this procedure would, in effect, weight corn doubly, once as grain and once as hogs. Likewise, an index number based on all production would include hay twice, once as hay and once as livestock and livestock products. The simplest way to eliminate such duplications in feed and livestock is to weight farm prices on the basis of sales.

The problem of duplication is common in other fields. For instance, index numbers of business activity frequently contain two series reflecting the same changes, such as steel production and carloadings.

It is not possible to generalize on the many other problems in choosing weights. Decisions must be made when the particular questions arise.

Variable Weights

Another aspect of the weighting problem has to do with the use of fixed or variable weights over long and short periods of time. It had been the common practice to use fixed weights, regardless of the length of the period covered by the index. In recent years, a few persons have varied the weights over long as well as short periods of time.

Changing the relative weights of commodities over a long period of time can be justified on the basis of upward or downward trends in relative production. For instance, during the last 200 years, the production of metals has increased relative to other things. New products have been introduced. For instance, petroleum was not discovered until about the middle of the nineteenth century, but since that time its production has increased at a very rapid rate. In addition, new uses have been found for old products. Cocoa is a product which was known and used for centuries, but was relatively unimportant until the chocolate bar appeared. Rubber is another product of this type. Some products formerly very important are now negligible. Ashes, candles, furs, and whiskey were once relatively much more important than at the present time.

Any change in weights should be slow and gradual in order that it have no effect on the short-time variations in the index of prices.

Changing weights over short periods of time usually has little justification. Trends in either prices or production cannot be determined at the time. The only possible basis for such changes in weights is year-to-year, month-to-month, or other short-time variations in production or sales. An index with short-time variable weights is not satisfactory because the weights are constantly changed among the different commodities so that the lowest-priced commodities get the greatest weight. Some students vary weights of indexes of farm prices in certain states from month to month with seasonal marketing. It is claimed that an index calculated with these weights shows the price situation as it affects producers in the particular month more clearly than an index with seasonally fixed weights. Variable seasonal weights allow the exclusion of perishable products from certain months in which they are not available or important. Seasonally changing weights may make comparisons between the corresponding months of different years more valuable, but they prevent the direct comparison of the farm price level for different months of the same or other years.

BASE PERIODS

One of the important requisites of a suitable base is its nearness to the period studied. Since the period studied is usually the present, the base should be well within the memory of most persons now doing business. Memory is short, and the human mind naturally makes comparisons most easily with events in the not-too-distant past.

Whenever possible, a base period in which the price structure was approximately in equilibrium should be chosen. Whenever prices rise or fall, some prices change more rapidly and by a greater amount than others. A disequilibrium in the price structure is then said to exist. Such a disequilibrium is greatest when prices are changing rapidly. Such a period should be avoided in the choice of a base, because:

(*a*) It is usually assumed that differences between commodities for any period are accurately shown by the individual indexes. This is to assume that a "normal" relationship existed during the base period; and

(*b*) In the absence of trend in relative prices, effective weights are most likely to coincide with given weights when the prices were in approximate equilibrium in the base period.

TYPE OF COMMODITIES

Volumes have been written on methods of calculation, weighting, base periods, and time- and factor-reversal tests for index numbers. These problems have been of considerable theoretical interest to a few students, but of much less practical importance. As pointed out in the above discussion, index numbers vary but little with method of calculation or weighting. Furthermore, they bear much the same relation to one another regardless of base period.

The major part of the variability in index numbers of prices is due to the commodities included and to the passage of time. Variability due to time is the primary purpose for which index numbers of prices are calculated and is usually taken for granted. Variability due to the commodities included has also been conspicuously absent in theoretical controversies around indexes of prices, either because it was not recognized, or because it also was taken for granted.

When the price level changes, some commodities change more rapidly and a greater amount than others. In 1932, when most prices were relatively low, farm prices were relatively lower than most other prices. Regardless of the method used and whether the commodities were weighted, the index of five farm products was about one-half that for five food products at retail (table 14).

TABLE 14.—EFFECT OF TYPE OF COMMODITY, WEIGHTING, AND METHOD ON INDEX NUMBERS FOR 1932

1910–1914 = 100

Type of commodity prices included	Indexes computed by various methods				
	Unweighted relatives		Weighted relatives		Aggregative
	Arithmetic mean	Geometric mean	Arithmetic mean	Geometric mean	
Farm, flexible:					
Corn, wheat, butter, hogs, cotton	52.9	51.2	51.6	50.3	51.8
Wholesale, flexible:					
Scrap steel, hides, lard, copper, coke	63.6	59.1	60.5	57.0	60.2
Wholesale, inflexible:					
Paper, rails, thread, sodium bicarbonate, cement	169.1	167.4	168.8	166.7	168.9
Retail, inflexible:					
Corn meal, hens, rib roast, milk, potatoes	117.3	116.9	117.6	117.2	117.5

Some wholesale prices declined very rapidly, while others changed but little. An index of scrap steel, hides, lard, copper, and coke was about one-third the index for paper, rails, thread, sodium bicarbonate, and cement (table 14). In 1932, the index for the latter group was about 170; and for the former, 55–60, when pre-war was 100. The size of an index of wholesale prices is, therefore, dependent on different combinations of the two types of commodities included. The commodities to be included in an index of prices can be classified in many ways. Some classifications give commodity groups whose price movements are widely divergent, such as the grouping into raw, semi-manufactured, and manufactured goods. In contrast, some classifications are made according to the flexibility of prices. Most classifications are merely descriptive of the commodities included, such as the division into metals, foods, and building materials.

During the past generation, there has been a trend toward further breaking down of price indexes into smaller groups. The motive for this has been fairly consistent in that the subgroups have been homogeneous as to type of commodity rather than as to type of price movement. In some cases, these classifications have incidentally resulted in homogeneity of price movement as well.

Since index numbers depend much more on the type of commodity included than upon the various statistical techniques, more emphasis should be placed on the former.

CHAPTER 5

SECULAR TREND

The tendency for things to vary is common and has been taken for granted by most persons in everyday life. The measurement of variability has already been discussed in some detail (chapter 3). A problem of far greater importance is the analysis of variability according to its causes. The variability in many types of data, such as prices, production, yields, and business activity, can be studied in reference to the passage of time. The relation of chronological differences to economic factors has long been studied and is one of the important problems in the field of economics.

Variations associated with the passage of time are of several types. For example, prices change from day to day, month to month, year to year, and from decade to decade. Some of these movements follow definite patterns, while others are quite irregular. One of the more persistent types of variation is secular trend. Secular trend is usually thought of as a persistent change occurring over a long period of time. In some analyses of any time series, it is often desirable to measure the amount of secular trend separately from other types of variability. In other analyses, it is desirable to eliminate secular trend from the data.

LINEAR TRENDS

Methods of Approximation

The most easily recognized type of secular trend is linear; that is, graphically it follows a straight line. Straight-line trend is easily understood because the rate of change is constant. Methods of calculating linear trends are relatively simple compared to methods of calculating non-linear trends.

Ruler or String Method

The simplest way to determine linear trend is to estimate it from the plotted data. After a little experience, this is a rather accurate method. It is widely used in preliminary analysis. There are certain differences in technique, such as superimposing a string or a transparent ruler over the plotted data. Usually, the line of trend is drawn with a straight edge after the position of the line has been established by inspection (figure 1). The line of trend for watermelon acreage was determined and drawn with the use of a transparent, celluloid ruler. By reading the values from this trend line for the first and last years, 1919 and 1937, it was found that the acreage increased from 142,000 to 280,000. The increase, 138,000 acres, spread over 18 years, amounted to 7,667 acres per year.

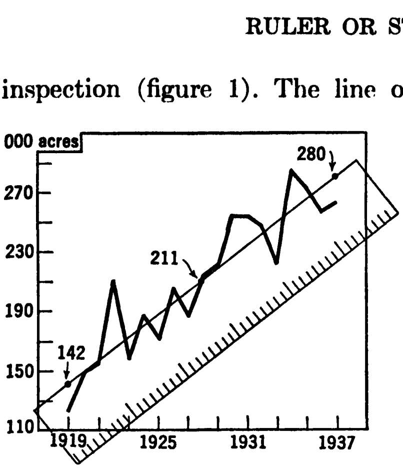


FIGURE 1.—RULER METHOD OF APPROXIMATING A STRAIGHT-LINE TREND

WATERMELON ACREAGE HARVESTED IN THE UNITED STATES, 1919–1937

The transparent ruler was shifted about until its upper edge appeared to represent the line of trend that approximately bisected the data.[1] The values for the first, middle, and last years were then read.

The average acreage for all 19 years can be approximated from the trend line. If the trend line has been drawn correctly, the average and the trend value for the middle year, 1928, will be identical. This value, 211,000 acres, is also the average of the values for the beginning and ending years [(142,000 + 280,000) ÷ 2 = 211,000].

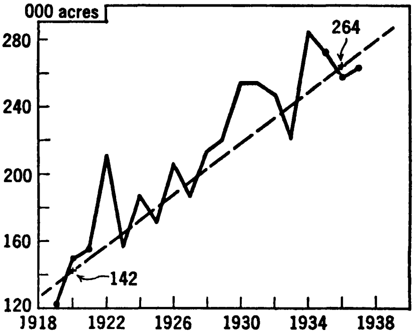


FIGURE 2.—AVERAGES METHOD OF APPROXIMATING A STRAIGHT-LINE TREND

WATERMELON ACREAGE HARVESTED IN THE UNITED STATES, 1919–1937

A straight line is drawn through the averages for the first 3 years, 142, and the last 3 years,[2] 264.

The yearly increase may be expressed as a percentage of the average for the period, the trend value for 1928. The annual increase was 7,667 acres, which is an annual increase[3] of 3.6 per cent (7,667 ÷ 211,000 × 100 = 3.6).

---

[1] A line that bisects the data assumes that the areas above and below the line are about equal. This is approximately, but not exactly, the same as the least-squares line about which the sum of the squared deviations is a minimum.

[2] From 1919 to 1921, the average acreage harvested was 142 (122, 149, and 155); and from 1935 to 1937, 264 (273, 257, and 263).

[3] Since the average and the annual increase are the same as the constants in an equation of a straight line, the equation may be written as follows:

$$Y = 211{,}000 + 7{,}667x.$$

Averages Method

Certain arbitrary guides for the estimation of the trend line have been used. Some workers designate two points representing the averages for a few years at each end of the data as determining the trend line (figure 2). In the watermelon illustration, the average for the first three years, centering on 1920, was 142,000; and for the last three years, centering on 1936, it was 264,000 acres. The difference, 122,000 acres, which was spread over a 16-year period, averaged 7,625 acres per year.[4] This rate of increase was about the same as that found in the ruler or string method of approximations.

Selected-Points Method

A variation of the averages method of approximation is that commonly designated as selected points. Two points determining the trend line are found by inspection rather than by averaging. The common practice is to use the values of normal or typical years. For instance, in the watermelon illustration, the years 1920 and 1935 appear to be approximately normal or typical of the years near the beginning and the end of the period studied. The acreages for these two years were 149,000 and 273,000. The increase of 124,000 acres in the 15-year period averaged 8,267 acres per year. If the years 1921 and 1936 had been selected as the normal, the increase would have been 102,000 (257,000 − 155,000 = 102,000), and the annual increase would have been 6,800 acres.

The accuracy of this method depends upon the skill with which the determining points are selected. The selected-points method of approximation is the least accurate one given because of the difficulty of choosing suitable points. The averages method is also inferior to the ruler or string method since the averages are based on limited data subject to chance variations. The ruler or string method is superior because the estimation of trend line is based upon all the data rather than a small portion of them.

Semi-Average Method

In this method, the straight line is based on two points determined by averaging the first and last halves of the data. The two averages are plotted at the center of their respective periods. The first 9-year average[5] for the watermelon acreage, 1919–1927, was 171,000; and the second, 1929–1937, was 253,000 acres. The first average was centered on 1923; and the second, on 1933. The annual increase was 8,200 acres per year (82,000 ÷ 10 = 8,200). This method is very simple, and the result does not depend on individual estimates. If the data are not subject to irregular and violent fluctuations, the method is reasonably satisfactory.

---

[4] If 4-year averages, 159,000 and 269,000 acres, were used, and the difference, 110,000, spread over 15 years, the annual rate of increase would be 7,333 acres. The two lines based on 3-year and 4-year averages would be different.

LEAST-SQUARES METHOD

Those students not satisfied with approximation methods may determine the best-fitting straight line by least squares. By this method, the average and the rate of increase are determined mathematically and are based on all the observations.

A straight line may be expressed algebraically as follows:

$$Y = a + bx$$

For annual data, it may be written diagrammatically as follows:

$$\begin{array}{c}\text{Estimated}\\ \text{yearly}\\ \text{values}\end{array} = \left(\begin{array}{c}\text{Average}\\ \text{value}\\ \text{for the}\\ \text{period}\end{array}\right) + \left(\begin{array}{c}\text{Average}\\ \text{yearly}\\ \text{rate of}\\ \text{change}\end{array}\right)\left(\begin{array}{c}\text{Years}\\ \text{measured}\\ \text{from middle}\\ \text{year}\end{array}\right)$$

The "average yearly rate of change" determines the slope of the line, while the size of the "average value for the period" determines its level.

In the equation $Y = a + bx$, Y is the trend value for the given year; a and b are constants; and x represents the passage of time. When the values of a and b are established, the value of Y in any trend line is dependent on variations in x, the passage of time.

The value of a, the average for the period, is calculated by the long-established method of summing the items in the series and dividing by the number of years.[6] The value of b, the rate of change or slope of the line, is less easily understood and is obtained from the following formula: $b = \Sigma xY \div \Sigma x^2$.

---

[5] The middle year, 1928, was omitted because the total period covered an odd number of years. The 9-year averages were calculated from the data in table 1, page 80.

[6] The general equation for a straight line in the slope-intercept form is $Y = a + bx$, in which b is the slope of the line, and a is the Y intercept.

In applying this formula to a line of secular trend, let the Y axis pass through the midpoint of the series of years on the X axis. Then the years will be denoted as -2, -1, 0, $+1$, $+2$, etc., either side of this point on the X axis. The Y axis will then bisect the trend line, and the value of the Y intercept will be the midpoint of this line, or the average of the series, i.e., $\Sigma Y \div N$. The slope is the increment in Y corresponding to a unit change in x. It can be demonstrated that the value of the slope is $\Sigma xY \div \Sigma x^2$.

TABLE 1.—CALCULATION OF STRAIGHT-LINE TREND BY METHOD OF LEAST SQUARES WITH ORIGIN AT THE MIDDLE OF AN ODD NUMBER OF YEARS

WATERMELON ACREAGE HARVESTED IN THE UNITED STATES, 1919–1937

Year	Deviation from middle year x	Deviations squared x^2	Acreage harvested, 000 omitted Y	Product of deviation in years and acreage xY	$Y = a + bx$ $a = \frac{\Sigma Y}{N}$
1919	−9	81	122	−1,098	$= \frac{4,031}{19} = 212.2$
1920	−8	64	149	−1,192	
1921	−7	49	155	−1,085	
1922	−6	36	211	−1,266	$b = \frac{\Sigma xY}{\Sigma x^2}$
1923	−5	25	158	− 790	
1924	−4	16	186	− 744	
1925	−3	9	171	− 513	$= \frac{4,342}{570} = 7.618$ ✓
1926	−2	4	205	− 410	
1927	−1	1	186	− 186	
1928	0	0	213	0	$Y = 212.2 + 7.618x$
1929	1	1	221	221	
1930	2	4	254	508	Per cent increase =
1931	3	9	254	762	
1932	4	16	248	992	
1933	5	25	221	1,105	$\frac{b}{a} \times 100$
1934	6	36	284	1,704	
1935	7	49	273	1,911	
1936	8	64	257	2,056	$= \frac{7.618}{212.2} \times 100$
1937	9	81	263	2,367	
Total	—	570	4,031	4,342	$= 3.59$

In fitting a straight line to watermelon acreage, the average acreage, a, is 212.2 (4,031 ÷ 19 = 212.2) (table 1). The yearly rate of change, b, is 7.618 (4,342 ÷ 570 = 7.618). The equation becomes:

$$Y = 212.2 + 7.618x$$

where 7.618 is the yearly increase in thousands of acres and 212.2 is the average acreage for the period 1919–1937 in thousands of acres.

The trend value for any particular year can be easily determined from the equation of the line by substituting for x in the equation the deviation corresponding to that year. For example, this deviation for 1934 was +6. The estimated or trend value was 258.

$$Y = 212.2 + (7.618)(+6)$$
$$Y = 212.2 + 45.708$$
$$Y = 257.9$$

The 1934 normal trend line value, 258, was somewhat less than the actual, 284.

The normal value for each year may be calculated and shown by dots lying in a straight line (figure 3). The usual practice is to calculate the values of only two dots and connect them with a straight line.

The average percentage increase[7] for the least-squares straight-line trend is given by dividing b by a and multiplying by 100. Stated another way, the slope of the line is expressed as a percentage of the average. The acreage harvested increased on the average 3.59 per cent per year $(7.618 \div 212.2 \times 100 = 3.59)$.

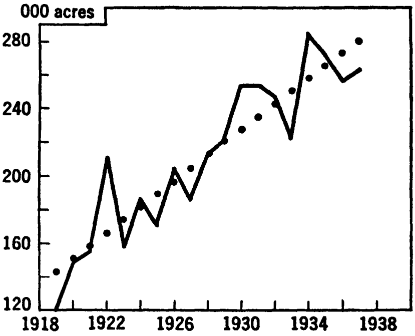


FIGURE 3.—LEAST-SQUARES METHOD OF CALCULATING A STRAIGHT-LINE TREND

WATERMELON ACREAGE HARVESTED IN THE UNITED STATES, 1919–1937

The points which form a straight line indicate the trend in the acreage. A common procedure is to indicate the trend with a continuous straight line.

One of the simplest techniques in determining the least-squares straight line is that used in table 1 when the origin was at the middle year and the number of years was odd. When the number of years is even, the problem is slightly complicated by the fact that there is no middle year. This difficulty is overcome by designating the point halfway between the two middle years as the origin of the deviations. Deviations from this origin are then expressed in half-years rather than years. In fitting a straight line to the acreage of watermelons from 1919 to 1938, a 20-year period, the origin was placed between 1928 and 1929. The deviations in terms of half-years were +1 for 1929 and +3 for 1930, and so on. The sums of the columns are $\Sigma 2x$, $\Sigma 2xY$, and $\Sigma 4x^2$, instead of Σx, ΣxY, and Σx^2, as they would be for an odd number of years. To obtain ΣxY, the quantity $\Sigma 2xY$ is divided by 2. Likewise,

---

[7] The average percentage rate of change is not to be confused with the constant percentage rate of change which could be calculated from the curve of compound interest type, $Y = ar^X$, where r is the rate.

$\Sigma 4x^2$ is divided by 4 to obtain Σx^2. From this point on, all calculations are the same as for an odd number of years as given in table 1.

The determination of the least-squares trend line is more difficult than the approximations by the various methods discussed. It is widely used and possesses the advantages of greater accuracy, rigidity of definition, and adaptability to further algebraic manipulation.[8] In practice, the most desirable method of determining the trend line depends upon the degree of accuracy desired. If there is to be no further algebraic treatment of the data, the approximate methods are usually about as satisfactory as the least-squares method. The trend lines for watermelon acreage by the different methods were similar.

Since the averages and yearly rates of change are known, they may be expressed in equation form, in thousands of acres, as follows:

String or ruler	$Y = 211.0 + 7.667x$
Averages	$Y = 203.0 + 7.625x$
Selected points	$Y = 215.1 + 8.267x$
Semi-average	$Y = 212.0 + 8.200x$
Least squares	$Y = 212.2 + 7.618x$.

The rates of change were quite similar. These trend lines were determined independently of one another and in the order given above. The beginning student cannot expect such a high degree of accuracy in estimation, but progress will be rapid.

The averages method places the trend line at a somewhat lower level than the other methods. This is indicated by the size of the first terms of the equations, which are the averages of all the points on the lines. By the least squares, this is also the average of the actual values. There is marked similarity in both the slope and level of the lines determined by the least-squares and ruler approximations.

NON-LINEAR TRENDS

In linear trends, rates of change are constant. Generally, however, the trends are not linear; that is, the rates are not constant because the factors responsible for the changes are themselves continually changing. Therefore, most secular trends are not linear. Nor do most trends follow any other definite pattern for a very long period of time. Many students have attempted to fit various types of mathematical curves to trends. They have been confronted with the difficulty of finding curves which fit the data and, more important, the particular curves which agree with the principle of the trend movement. Usually, they have failed to find rigid curves which satisfy these requirements. They have also encountered the practical difficulty of the great amount of work which the calculation of these curves usually involves.

---

[8] The significance of this trend line is easily and accurately tested.

AN EXPONENTIAL OR COMPOUND-INTEREST CURVE

Some phenomena increase at a uniform, proportionate rate throughout the series. It happens that there is a particular type of curve that expresses this same principle. This type of exponential curve, the compound-interest curve, is given by the equation $Y = ar^X$ and is not too difficult to calculate.[9] From 1919 to 1938, the production of grapefruit increased at the uniform, proportionate rate of 9.95 per cent per year (figure 4).

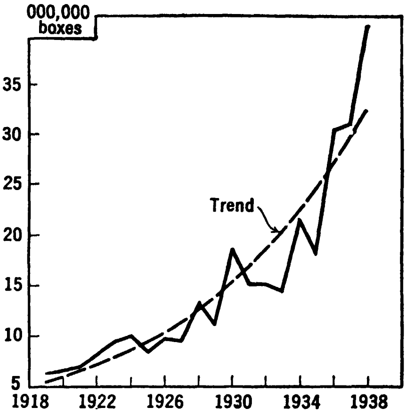


FIGURE 4.—LEAST-SQUARES METHOD OF CALCULATING AN EXPONENTIAL CURVILINEAR TREND

PRODUCTION OF GRAPEFRUIT IN 4 STATES, 1919–1938

The trend in the production of grapefruit increased at a compound rate of 9.95 per cent per year.[10]

MOVING AVERAGES

There is a wide demand for a method of measuring trend in which the calculations are relatively simple and the lines or curves are sufficiently flexible to fit a wide variety of data. Moving averages are the most widely used tools to describe non-linear trends. Moving averages are series of arithmetic means of a variable for a given number of units of time. As time passes, the values for earlier periods are replaced in the means by values for succeeding periods. The whole series of successive arithmetic means is termed moving averages.

The Washington production of apples increased at a rather rapid rate until about 1925, and since then has leveled off (figure 5). The moving-average method is well adapted to this type of trend.

Constructing the moving average is not difficult but involves considerable calculation. The sum of the production figures for the 7 years

---

[9] By the least-squares method, the constants a and r for the compound-interest curve $Y = ar^X$ are determined by solving the following simultaneous equations:

$$N \log a + \log r \Sigma X - \Sigma(\log Y) = 0$$
$$\log a \Sigma X + \log r \Sigma X^2 - \Sigma(X \log Y) = 0$$

[10] $Y = 5.3803\ (1.0995)^x$.

1905 to 1911 is the first 7-year moving total. This total, 24.5, is placed opposite the seventh year, 1911, merely because it is the simplest and least confusing mechanical procedure (table 2). The first 7-year moving average is this moving total, 24.5, divided by 7, or 3.5. Since it is centered on the fourth year of the seven, the moving average, 3.5, is placed opposite the year 1908.[11] The second moving totals and averages are 29.7 and 4.2, respectively. The same procedure is followed throughout the series. These moving totals may be obtained by adding each 7-year period independently. They may also be obtained by subtracting from the previous 7-year total the first item of the seven and then adding the next item after the seventh. For instance, the first 7-year total based on 1905 to 1911 was 24.5. The next moving total is based on 1906 to 1912 and may be derived by subtracting 2.5 from and adding 7.7 to the 1905–1911 total, 24.5. The result is 29.7 (24.5 − 2.5 + 7.7 = 29.7).

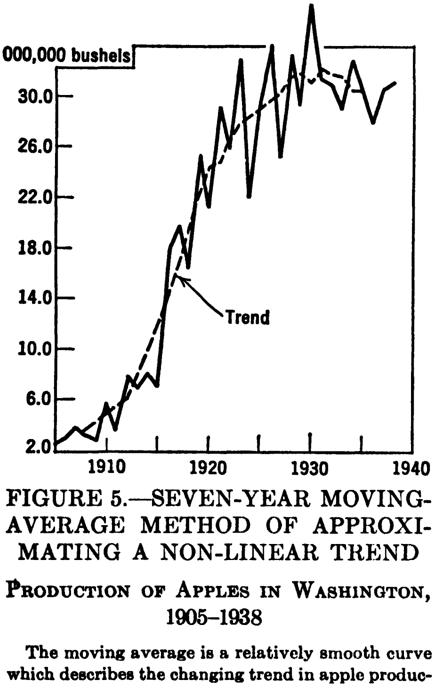


FIGURE 5.—SEVEN-YEAR MOVING-AVERAGE METHOD OF APPROXIMATING A NON-LINEAR TREND

PRODUCTION OF APPLES IN WASHINGTON, 1905–1938

The moving average is a relatively smooth curve which describes the changing trend in apple production.

One of the above procedures is followed until the work is complete. The first procedure has the advantage that each moving total is an independent calculation, and errors are not cumulative. The second method requires less time, but errors in any one 7-year period carry over into all the following periods. When it is used, the computations should be checked at intervals by summing the items in the moving total. When only three or four units of time are included in the moving average, the first procedure is preferable; when seven or more, the second procedure is preferable.

---

[11] The general tendency is to place moving averages opposite the middle year. A 5-year moving average is centered on the middle or third year; and a 10-year average, on the fifth or sixth year. There are many variations in moving-average technique which place the average on any one of the years included, or on the year following.

TABLE 2.—CALCULATION OF TREND BY A 7-YEAR MOVING AVERAGE

PRODUCTION OF APPLES IN WASHINGTON, 1905–1938, IN MILLIONS OF BUSHELS

Year	Production	Seven-year moving		Year	Production	Seven-year moving	
		Total	Average			Total	Average
1905	2.5	—	—	1922	25.8	155.7	26.6
1906	3.0	—	—	1923	33.0	171.0	27.9
1907	3.8	—	—	1924	22.0	173.2	28.4
1908	3.2	—	3.5	1925	29.6	186.3	29.0
1909	2.7	—	↗ 4.2	1926	34.0	195.0	29.6
1910	5.8	—	4.8	1927	25.3	198.8	30.3
1911	3.5	24.5	5.4	1928	33.5	203.2	31.6
1912	7.7	29.7	6.0	1929	29.5	206.9	31.8
1913	6.9	33.6	8.2	1930	37.9	211.8	31.1
1914	8.3	38.1	10.2	1931	31.4	221.2	32.2
1915	7.3	42.2	12.0	1932	31.0	222.6	31.8
1916	17.7	57.2	14.5	1933	29.2	217.8	31.6
1917	19.8	71.2	16.6	1934	33.0	225.5	30.5
1918	16.5	84.2	19.6	1935	30.7	222.7	30.5
1919	25.3	101.8	22.2	1936	28.0	221.2	↗ —
1920	21.5	116.4	24.4	1937	30.5	213.8	—
1921	29.1	137.2	24.7	1938	31.1	213.5	—

The 7-year moving averages of apple production in Washington given in table 2 are shown graphically in figure 5. Production was subject to violent year-to-year fluctuations due to yield, and to long-time fluctuations due to changes in acreage and age of trees. The 7-year moving average is a rather smooth line which describes the long-time changes, and the effect of yearly fluctuations is almost eliminated.

One of the most important problems in the use of the moving average is its length. It is desirable that the trend line be an approximately smooth line. Smoothness depends on the length of time covered by the moving average, the violence of short-time fluctuations in the data, and the length of these fluctuations. In general, the shortest moving average which will result in a reasonably smooth line is best. In deciding on the length, the short-time fluctuations must be examined in detail. In series describing the production of farm products, fluctuations are relatively violent, but are usually only one or two years in length. The moving average of apple production included 7 years. For crops with greater violence in production, a longer average might be desirable. It is also conceivable that for crops with less fluctuation a shorter average would be satisfactory. Some series, for example the number of hogs on farms, are subject to fluctuations several years in length. With a given degree of fluctuation, the length of the moving average necessary to iron out short-time changes increases with the length of these changes.

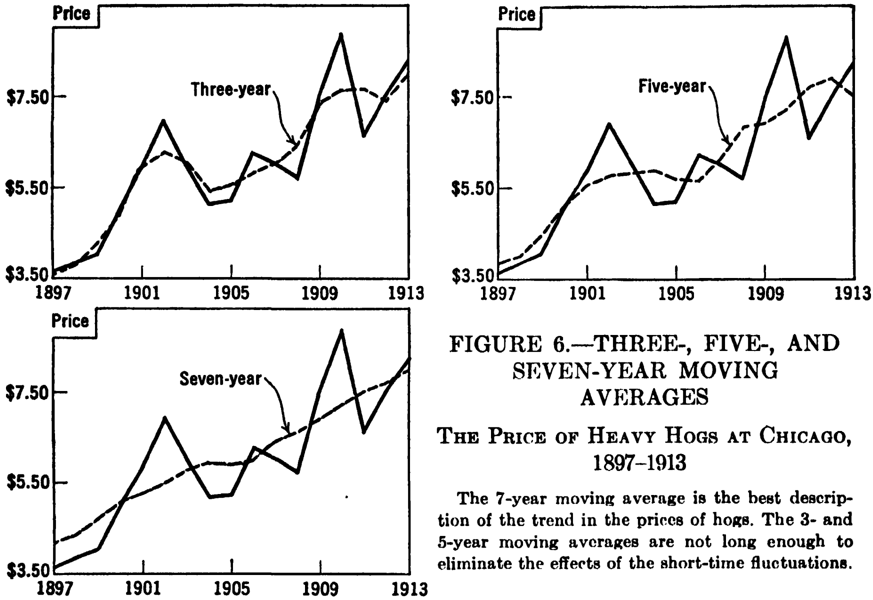


FIGURE 6.—THREE-, FIVE-, AND SEVEN-YEAR MOVING AVERAGES

THE PRICE OF HEAVY HOGS AT CHICAGO, 1897–1913

The 7-year moving average is the best description of the trend in the prices of hogs. The 3- and 5-year moving averages are not long enough to eliminate the effects of the short-time fluctuations.

A 3-year moving average of the price of hogs at Chicago was too short to give a satisfactory trend line (figure 6). It did not iron out the short-time changes. In fact, the 3-year moving average resembled the prices themselves. The high and low points of the series are so far apart that the 3-year moving averages included alternately 3 high years and 3 low years. As a result, they form an irregular trend line. When a 5-year moving average is used, the trend line is somewhat improved, but still contains in a lesser degree the irregularities of the 3-year average (figure 6). The 7-year moving average is almost a straight line and shows the trend and nothing else. Seven years is a long enough period to include counterbalancing high and low years.

Generally, moving averages deviate somewhat from a smooth curve. In fact, it is rarely possible to construct a smooth curve by moving averages. Since extending the length of the moving average tends to increase smoothness, the correct length of average to use depends on the degree of smoothness desired. With this degree of smoothness in mind, the student usually determines the length by trial and error. In general, this desired length increases with (*a*) the length of short-time fluctuations, (*b*) the irregularity of this length, (*c*) the violence of the fluctuations, and (*d*) the irregularity of this violence.

It is not possible to calculate a moving average for every item in the series. In a 7-year moving average, centered on the fourth year, there would be no moving averages for the first 3 or the last 3 years (table 2). This loss of data at the ends of the series is sometimes a serious disadvantage of the moving-average method. Some students have attempted to remedy this difficulty by arbitrarily extending the moving average through to the ends of the series. A common procedure is to project the trend line in the direction indicated by moving averages near the ends. These efforts are usually based on guesswork and are often in error. Nevertheless, these extensions are frequently as reliable as the values at the ends of a mathematically determined line.

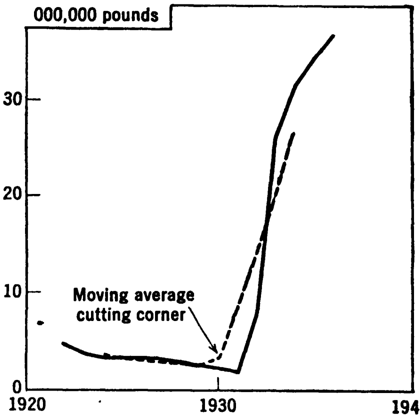


FIGURE 7.—MOVING AVERAGES "CUT CORNERS"

BREWERY CONSUMPTION OF HOPS, 1922–1936

During 1932–1933, there was a very sharp reversal of the downward trend in hop consumption. The upturn in the 5-year moving average precedes the actual change by one to two years, thereby "cutting the corner."

A commonly cited fault of moving averages is their tendency to "cut corners." Moving averages do not follow non-linear data to the highest and lowest points. The very quality of the moving average which makes it useful in smoothing a curve is disadvantageous when there are sharp turns in the trend.

During the twenties, the consumption of hops was low and declined slowly. With the end of prohibition in the early thirties, consumption increased very rapidly (figure 7).

CROP YEAR	BREWERY CONSUMPTION, MILLION POUNDS	FIVE-YEAR MOVING AVERAGE
1929	2.6	2.5
1930	2.2	3.4
1931	1.8	8.1
1932	7.8	14.0
1933	26.2	20.4

The first sharp increase in consumption came in the crop year 1932; and the greatest increase, in the following year, 1933. The 5-year moving average started up as early as 1930, and a sharp increase took place in 1931. Thus, this moving average does not present the true picture of the upward trend which actually began later and was steeper than indicated.

Methods[12] have been devised to adjust moving averages to compensate for their tendency to "cut corners." Cutting corners is not confined to moving averages. Regardless of the type of trend line used, the problem of cutting corners arises whenever there are abrupt changes in trend.

There has been considerable controversy over the reliability of moving averages for indicating trend. The shape of the moving-averages trend is always determined entirely by the data themselves. The length of the moving average affects the flexibility of the curve. The effect of short-time variations is never completely removed. In general, moving averages are more flexible than mathematical curves.

A mathematically fitted trend depends partly on the data, but is limited to a rigidly smooth curve. Moreover, any mathematical curve follows some definite pattern. This allows the student to exercise some rigidity in determining a line conforming with the principles behind the trend movement. However, with a few exceptions, these principles are unknown, and the above advantage is of little practical value.

USES

One of the uses of trends is the comparison of the rates of change in various types of prices, production, and other economic phenomena during the same or different periods of time. From 1839 to 1914, the total basic production of the United States increased 4.03 per cent per year (table 3). The equation for this growth was: $Y = 4.14(1.0403)^X$. The rate of increase, 4.03, which is a constant proportion, is read from that part of the equation in the parentheses. The first term is the normal value of the index for the first year, 1839, when 1926–1930 = 100. Since basic production increased more rapidly than population, 4.03 compared with 2.28, each individual produced and presumably consumed more product with passing time. The ratio of these two rates, 1.7, measures the improved standard of living of the American people ($1.0403 \div 1.0228 = 1.017$). Urban activity, as measured by the production of fuel and power, and other minerals and secondary metals, increased more rapidly than agricultural production (table 3).

---

[12] Brandow, G. E., Cycles in Industry and Prices, Appendix A, p. 14, 1939. Unpublished manuscript, Cornell University.

Enström, A. F., On Periodicities in Climatic and Economic Phenomena and Their Covariation, Ingeniörsvetenskapsakademien, Handlingar, Nr. 31, Stockholm, 1924.

TABLE 3.—RATES OF CHANGE* IN POPULATION AND PRODUCTION IN THE UNITED STATES,† 1839–1914 AND 1915–1929

Index of	Rate for period	
	1839–1914	1915–1929
Population	2.28	1.46
Crop production	3.03	0.85
Fuel and power	5.96	4.84
Minerals and secondary metals	7.02	3.62
Total basic production	4.03	2.11

* Based on $Y = ar^X$.

† Warren, G. F., and Pearson, F. A., The Physical Volume of Production in the United States, Cornell University Agricultural Experiment Station Memoir 144, p. 7, November 1932.

The above comparisons are based on one period of time. A second type of comparison may be made between the trends in different periods. For instance, from 1915 to 1929, the rate of change in population, 1.46, was much less than that for 1839–1914, 2.28. During the latter period, most types of production experienced diminishing rates of increase. The most notable was in agriculture. The annual rate of change in agricultural production declined from 3.03 to 0.85 per cent per year.

Another type of comparison is the change in two rates during two periods of time. In the first period, crop production rose more rapidly than population. After 1914, it did not keep pace with population, with the result that exports of agricultural products decreased rapidly. If agricultural production continues to increase less rapidly than population, exports will decline and be replaced by imports.

One of the most important uses of trends comes in further analysis of time series. In the study of cycles, supply-price relationships, and the like, the long-time trend must be eliminated before the effect of other factors can be studied. The problem of eliminating trend is discussed in chapters 6 and 7.

CHAPTER 6

SEASONAL VARIATION

Nearly all products vary in demand with the seasons of the year. Many products are of necessity produced during only a part of the year. The seasonal variation for production may be the same as for demand, or quite different. For example, coal is in greatest demand in the winter, and production is greatest at that time. Eggs are most desired in cold weather, but hens lay nearly half the yearly production in four spring and summer months.

Because of seasonal variation in demand and in production, there is also seasonal variation in prices and market movements.

Since manufacturing can be adjusted to demands more easily than agriculture, seasonal variation is a less important problem in most types of industry than in agriculture.

The error is frequently made of comparing prices for a given month with a yearly average price or with prices for some other month in order to determine whether the prices are high or low. Such comparisons frequently lead to erroneous conclusions. For this reason, it is desirable that one know the normal seasonal variation of prices, production, distribution, and the many other activities of our daily life.

SIMPLE AVERAGES METHOD

The easiest and one of the more common methods of measuring seasonal variation is to average the data for each month for a series of years. For example, the average January price of heavy hogs at Chicago from 1897 to 1913 was $5.65 (table 1). The corresponding average for April was $6.31. The average for the period was $6.03. The average January price of hogs was 93.7 per cent of the average for the entire period ($5.65 \div 6.03 = 0.937$). The corresponding index for April was 104.6.

If there is no pronounced secular trend in the series, this is the simplest and a reasonably satisfactory method of calculating an index of seasonal variation. It is widely used because of its simplicity. Errors resulting from its use are ordinarily small, but sometimes are large enough to lead to erroneous conclusions.

TABLE 1.—SIMPLE AVERAGES METHOD OF CALCULATING SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Dollars per 100 lb.

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Average
1897	3.35	3.35	3.85	4.05	3.75	3.40	3.50	3.90	4.00	3.75	3.40	3.35	—
1898	3.65	4.00	3.90	3.90	4.35	4.10	3.95	3.90	3.85	3.70	3.45	3.40	—
1899	3.75	3.80	3.80	3.85	3.90	3.80	4.25	4.55	4.40	4.30	3.90	4.05	—
1900	4.55	4.90	5.00	5.55	5.30	5.20	5.25	5.20	5.25	4.80	4.80	4.75	—
1901	5.25	5.40	5.90	5.85	5.80	6.00	5.90	5.95	6.65	6.10	5.70	6.20	—
1902	6.40	6.30	6.50	7.10	7.00	7.50	7.80	7.25	7.55	7.00	6.35	6.35	—
1903	6.60	7.00	7.45	7.30	6.60	6.05	5.45	5.30	5.75	5.40	4.60	4.50	—
1904	4.95	5.25	5.50	5.15	4.75	5.30	5.35	5.25	5.70	5.35	4.80	4.50	—
1905	4.70	4.90	5.20	5.45	5.40	5.30	5.60	5.90	5.40	5.10	4.80	4.90	—
1906	5.40	6.00	6.30	6.50	6.45	6.55	6.60	6.15	6.15	6.40	6.20	6.25	—
1907	6.60	7.05	6.65	6.60	6.35	6.05	5.90	5.90	5.80	6.05	4.90	4.65	—
1908	4.45	4.50	5.05	5.85	5.50	5.80	6.55	6.60	6.90	6.05	5.90	5.75	—
1909	6.20	6.45	6.80	7.30	7.40	7.80	7.90	7.60	8.10	7.85	8.10	8.45	—
1910	8.70	9.20	10.65	10.00	9.50	9.35	8.60	8.25	8.70	8.45	7.55	7.65	—
1911	7.85	7.25	6.70	6.15	5.85	6.15	6.65	7.15	6.75	6.50	6.35	6.25	—
1912	6.30	6.25	7.10	7.85	7.70	7.50	7.60	8.05	8.30	8.65	7.75	7.45	—
1913	7 40	8.05	8.75	8.80	8.40	8.50	8.95	8.10	8.10	8.15	7.80	7.70	—
Total	96.10	99.65	105.10	107.25	104.00	104.35	105.80	105.00	107.35	103.60	96.35	96.15	—
Average	5.65	5.86	6.18	6.31	6.12	6.14	6.22	6.18	6.31	6.09	5.67	5.66	6.03
Index	94	97	102	105	101	102	103	102	105	101	94	94	100

TREND-ADJUSTED METHOD

If there is any secular trend in the data, the simple average method gives incorrect results. During the period 1897–1913, the prices of hogs at Chicago were generally rising. For this reason, the December prices, which are 11 months later than the previous January prices, would tend to be somewhat higher. Similarly, November prices would average higher than those for February. This would tend to make the index of seasonal variation low in the first half of the year and high in the second half. This difficulty has resulted in many methods of correcting seasonal indexes for trend. One of the simplest methods is illustrated below.

During the period 1897–1913, the equation of the secular trend of the price of hogs was: $Y = \$6.034 + \$0.248x$. The price of hogs increased $0.248 per year, $0.02067 per month, or $0.01033 per half-month. Correction may be made by taking the middle of the year as a base and adding or subtracting each way (table 2). For example, add half a month's correction to June and deduct the same amount from July; subtract one and a half months' from August, $0.03, and add the same amount to May. This procedure is continued until 11 half-month intervals are deducted from December and added to January.

The corrected January price was $5.76, based on the average January price, $5.65, plus the 11-cent correction. The average of the 12 monthly corrected prices, $6.03, is, of course, equal to the average of the original prices. The index number of seasonal variation is obtained by dividing the monthly corrected prices by the average, $6.03. The index of the January price was 95 (table 2).

TABLE 2.—TREND-ADJUSTED METHOD OF CALCULATING SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Month	Average price*	Half-month interval	Correction†	Corrected price	Seasonal index
January	$ 5.65	11	$ +0.11	$ 5.76	95
February	5.86	9	+0.09	5.95	99
March	6.18	7	+0.07	6.25	104
April	6.31	5	+0.05	6.36	105
May	6.12	3	+0.03	6.15	102
June	6.14	1	+0.01	6.15	102
July	6.22	1	−0.01	6.21	103
August	6.18	3	−0.03	6.15	102
September	6.31	5	−0.05	6.26	104
October	6.09	7	−0.07	6.02	100
November	5.67	9	−0.09	5.58	92
December	5.66	11	−0.11	5.55	92
Total	72.39	—	—	72.39	1,200
Average	6.03	—	—	6.03	100

* Table 1.

† The yearly equation of secular trend, 1897–1913, was: $Y = \$6.034 + \$0.248x$. The monthly increase was $0.02067; and for one half-month, the change was $0.01033. The corrections were as follows:

1 half-month	$0.010	5 half-months	$0.052	9 half-months	$0.093
3 half-months	0.031	7 half-months	0.072	11 half-months	0.114

If the secular trend is downward, the method of procedure is the same, except that the corrections are added to the last half of the year and deducted from the first.

When the secular trend is linear, the trend-adjusted method is very satisfactory. When there is a pronounced non-linear trend, other methods which are considerably more involved may give more satisfactory results.

MOVING-AVERAGE METHOD

By this method, the secular trend is removed by expressing the price for each month as a percentage of a moving average. The median of the ratios for each month is determined, and these 12 medians are adjusted so that they average 100.

TABLE 3.—MOVING-AVERAGE METHOD OF CALCULATING SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Month	Price*	Moving average†	Ratio	Month	Price*	Moving average†	Ratio
1897				.	.	.	.
January	$3.35	—	—	.	.	.	.
February	3.35	—	—	.	.	.	.
March	3.85	—	—	1912			
April	4.05	—	—	July	$7.60	$7.54	101
May	3.75	—	—	August	8.05	7.63	106
June	3.40	—	—	September	8.30	7.78	107
July	3.50	$3.64	96	October	8.65	7.92	109
August	3.90	3.66	107	November	7.75	8.00	97
September	4.00	3.72	108	December	7.45	8.06	92
October	3.75	3.72	101	1913			
November	3.40	3.71	92	January	7.40	8.14	91
December	3.35	3.76	89	February	8.05	8.25	98
1898				March	8.75	8.26	106
January	3.65	3.82	96	April	8.80	8.24	107
February	4.00	3.85	104	May	8.40	8.20	102
March	3.90	3.85	101	June	8.50	8.20	104
April	3.90	3.84	102	July	8.95	8.23	109
May	4.35	3.84	113	August	8.10	—	—
June	4.10	3.84	107	September	8.10	—	—
. ‡	.	.	.	October	8.15	—	—
.	.	.	.	November	7.80	—	—
.	.	.	.	December	7.70	—	—

---

* Table 1.

† Twelve-month moving average centered on seventh month.

‡ The calculations from June 1898 to July 1912 were omitted to save space. All the ratios are given in table 4.

For hog prices, a 12-month moving average was calculated from 1897 to 1913 (table 3). For July 1897, the moving average was $3.64, while the actual price was $3.50. The ratio of the actual price to the moving average, 96, indicated the magnitude of the July price relative to the average price for the year in which July is centered (table 3). Similarly, the ratio of August 1897 was 107. This procedure is followed throughout the period. In using a 12-month moving average, a half-year is lost at each end of the data.

TABLE 4.—MOVING-AVERAGE METHOD OF CALCULATING SEASONAL VARIATION, CONTINUED

RATIOS OF THE WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO TO THEIR MOVING AVERAGE, 1897–1913

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
106	111	120	114	113	109	114	112	114	110	97	99
103	104	113	114	108	107	113	112	112	109	96	98
100	104	108	112	107	107	109*	111	110	105	96	98
99	104	108	110	106	107	107	110	108	102	96	97
99	103	107	110	105	104	106	107	108	101	94	93
98	103	106	107	105	104	105	106	107	101	94	92
98	102	105	105	104	104	105	104	105	100	93	90
96	100	104	105	104	104	104	102	105	100	92	90
96	100	101	105	102	103	*104*	102	103	99	92	90
95	100	101	105	102	103	103	101	102	99	91	89
94	98	99	105	102	102	101	101	102	98	90	89
93	96	99	105	101	101	100	99	101	97	90	89
92	95	99	102	99	98	100	99	101	97	89	87
91	95	97	100	99	98	98	97	101	95	88	87
91	90	93	99	92	95	97	94	100	93	86	86
83	83	92	88	85	91	96*	90	96	92	86	86
						91					
Median† 96.0	100.0	102.5	105.0	103.0	103.5	104.0	102.0	104.0	99.5	92.0	90.0
Index‡ 96	100	102	105	103	103	104	102	104	99	92	90

* Noted in text.

† The 12 medians totaled 1,201.5 and averaged 100.1.

‡ The 12 seasonal indexes averaged 100.

All the ratios for July were then arranged in order of size, and their median determined. In like manner, the medians were determined for other months. The ratios for July 1897 and July 1913 were 96 and 109 and were arranged with the other July ratios according to size (marked with asterisks in table 4). The median or middle item was 104.0. The remaining months were treated in the same manner. The sum of the 12 medians was 1,201.5; and the average, 100.1. The adjusted medians, obtained by dividing each month by 100.1, were the indexes of seasonal variation.

There are many variations of this method. Some use a longer moving average, and others calculate the arithmetic average of the ratios instead of determining the median. Still others average the three, four, or five middle ratios.

This method requires more computation than the simple average or trend-adjusted methods. It has the advantage that it is more flexible because it eliminates non-linear as well as linear trends.

LINK-RELATIVE METHOD

By this method, the value for each month is expressed as a percentage of that for the preceding month. The February 1897 price of hogs, $3.35, was expressed as a percentage of the January 1897 price, $3.35 (table 1, page 91). Therefore, the link relative was 100. The link relative for February 1898 was 110 (4.00 ÷ 3.65 = 1.10). The link relative for February in terms of the corresponding January was determined for all years.

The link relative for March 1897 was 115 (3.85 ÷ 3.35 = 1.15). The same procedure is followed to obtain the link relatives for the other months in all the years.

The next step is the arrangement of the link relatives according to magnitude to determine the medians for each month (table 5). The median link relative for January was 105. This is not a seasonal index. It is merely the median of the 16 ratios of each January divided by the previous December. However, a seasonal index may be constructed from these medians of link relatives. As a starting point, January was given a "converted value" of 100. Since the median link relative for February was 104, the converted value for February was also 104 (104 × 1.00 = 104). Since the median for March was 105, March had a converted value of 109.2 (104 × 1.05 = 109.2). The median for April, 101, was multiplied by the converted value for March, 109.2, and the product, 110.3, was the converted value for April. This is continued for the remaining months.

Since the medians of the link relatives are the average of each month's prices expressed as a percentage of the preceding month, the multiplication process to establish the converted values merely restores the approximate seasonal variation which was lost by the division in determining the link relatives. The result is an index comparable to the original values. These converted values, or chain relatives as they are sometimes called, were still not completely adjusted for trend. To establish the amount of adjustment necessary to correct for this trend and for peculiarities in the process, the converted value for January based on December was calculated. The January median, 105, multiplied by the converted December value, 94.4, was 99.1. This was not quite the same as the arbitrary value given to January, 100. The converted values for each month were adjusted so that the calculated converted value for January was also 100. The difference between the

TABLE 5.—LINK-RELATIVE METHOD OF CALCULATING SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Jan.	Av.
Link relatives													
112	111	116	116	112	112	113	111	112	104	103	109	—	—
111	110*	115*	111	101	107	112	108	109	104	100	104	—	—
110	109	114	111	101	105	108	107	108	104	98	104	—	—
110	108	112	109	99	105	106	106	107	101	98	102	—	—
110	107	109	107	99	105	105	105	105	98	97	101	—	—
109	106	109	105	99	103	104	101	105	97	96	101	—	—
108	106	106	105	99	102	103	101	104	97	94	100	—	—
106	106	106	103	98	101	101	100	103	96	93	99	—	—
104	104	105	101	96	98	101	99	103	96	93	99	—	—
104	104	105	101	95	98	101	99	101	94	91	99	—	—
103	103	105	100	95	98	101	98	100	94	91	99	—	—
103	101	103	99	95	97	101	97	100	94	91	98	—	—
103	101	102	99	95	97	98	96	99	94	90	98	—	—
101	100	100	98	94	95	98	96	98	93	90	97	—	—
99	99	98	94	93	94	96	93	97	92	89	96	—	—
96	98	94	94	92	92	92	93	94	91	85	95	—	—
	92	92	92	90	91	90	91	92	88	81	94	—	—
Median													
105	104	105	101	96	98	101	99	103	96	93	99	—	—
Converted values†													
100	104	109.2	110.3	105.9	103.8	104.8	103.8	106.9	102.6	95.4	94.4	99.1	—
Adjustment factor‡													
0	+0.1	+0.2	+0.2	+0.3	+0.4	+0.5	+0.5	+0.6	+0.7	+0.8	+0.8	+0.9	—
Adjusted													
100	104.1	109.4	110.5	106.2	104.2	105.3	104.3	107.5	103.3	96.2	95.2	—	103.9
Index													
96	100	105	106	102	100	101	100	103	99	93	92	—	—

* Link relatives computed in the text. † Sometimes called chain relatives.

‡ (January arbitrary value) − (January converted value) = 100 − 99.1 = 0.9; Monthly differential = 0.9 ÷ 12 = 0.075.

0×0.075 = 0	3×0.075 = 0.2	6×0.075 = 0.5	9×0.075 = 0.7
1×0.075 = 0.1	4×0.075 = 0.3	7×0.075 = 0.5	10×0.075 = 0.8
2×0.075 = 0.2	5×0.075 = 0.4	8×0.075 = 0.6	11×0.075 = 0.8

arbitrary and converted values for January, 0.9, was equal to a monthly differential of 0.075. This differential, multiplied by 1 for February, 3 for April, and 11 for December, gave the adjustment factors 0.1, 0.2, and 0.8, respectively. The adjustment factors for each month were added to the corresponding converted values to obtain the adjusted values.[1] For example, to the converted value for April, 110.3, was added the adjustment factor, 0.2, and the sum, 110.5, is the adjusted value.

---

[1] If the calculated value for January was greater than 100, the arbitrary value for January, the adjustment factors would be subtracted from, rather than added to, the monthly corrected values.

Some students use a geometric rather than arithmetic principle of adjustment.

These values were still not the final indexes of seasonal variation because they did not average 100. The adjusted value for each month was corrected by dividing by the average, 103.9, to obtain the final index. The index for April, 106, was obtained by dividing 110.5 by 103.9, and rounding to the nearest per cent (table 5).

The link-relative method is more difficult to understand than the other methods presented. The procedure involves two major parts, the calculation of the link relatives and the elimination of trend. The first part is rather laborious but minimizes the effects which cycles or nonlinear trends might have on the seasonal index. The second part, the elimination of trend and final adjustments, is rather difficult to comprehend but simple to calculate. The method has the merit that nothing is left to judgment.

COMPARISON

The four methods here discussed are quite widely used and all have been both attacked and defended. Space does not permit a summary of this controversy. In general, the link-relative and moving-average methods involve the most calculation but are the most flexible.

Many other methods of measuring seasonal variation have been developed but are not discussed here. Indexes of seasonal variation of the price of heavy hogs at Chicago were calculated by the Bauman moving-average-difference, the Carmichael first-difference, and the Falkner per-cent-of-trend methods and compared with the four methods discussed (table 6). All indexes indicate that hogs were high-priced during spring and summer and low in the winter. They all show that hog prices had seasonal peaks in April and September. Five of the seven indexes show that the December price was the lowest. In general, the simple average indexes for the first three months tended to be lower, and for the last three months higher, than the indexes obtained by the other six methods. Differences among the six indexes with the trend removed were generally small.

The period 1897–1913 was one in which prices generally rose at a rather uniform rate, and the different methods of calculating seasonal variation gave substantially the same results (table 6). The question may be raised concerning the relative accuracy of the various methods during a period of unusual price changes. The 11-year period 1928–1938 was one of violent fluctuations in all prices. The period was marked by deflation from 1929 to 1932, revaluation of the dollar in 1933, a rapid rise in 1936–1937, and a sharp decline in 1937–1938. Four methods were used to determine the seasonal variation of the price of hogs during this period. All methods indicate that peaks in prices occurred

TABLE 6.—COMPARISON OF SEASONAL INDEXES CALCULATED BY SEVEN METHODS

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Month	Simple average*	Trend adjusted†	Moving average‡	Link relatives§	Moving average difference‖	First difference¶	Per cent of trend**
January	94	95	96	96	97	96	94
February	97	99	100	100	101	99	100
March	102	104	102	105	106	104	102
April	105	105	105	106	107	106	107
May	101	102	103	102	102	103	106
June	102	102	103	100	100	102	102
July	103	103	104	101	102	103	100
August	102	102	102	100	101	102	101
September	105	104	104	103	103	104	104
October	101	100	99	99	99	99	99
November	94	92	92	93	92	92	92
December	94	92	90	92	91	91	91

* Table 1, page 91. † Table 2, page 92.

‡ Table 4, page 94. Macaulay method. Index of Production in Selected Basic Industries, Federal Reserve Bulletin, Vol. 8, No. 12, pp. 1416–17, December 1922.

§ Table 5, page 96. Persons, W. M., Indices of Business Conditions, The Review of Economic Statistics, Preliminary Volume, No. 1, p. 37, January 1919.

‖ Bauman, A. O., Thirteen-Months-Ratio-First-Difference Method of Measuring Seasonal Variation, Journal of the American Statistical Association, Vol. 23, New Series, No. 163, pp. 282–290, September 1928.

¶ Carmichael, F. L., Methods of Computing Seasonal Indexes: Constant and Progressive, Journal of the American Statistical Association, Vol. XXII, New Series, No. 159, pp. 339–354, September 1927.

** Falkner, H. D., The Measurement of Seasonal Variation, Journal of the American Statistical Association, Vol. XIX, New Series, No. 146, pp. 167–179, June 1924.

in the early spring and in the late summer (table 7). These results[2] were approximately the same as for the previous period (table 6). The results by the various methods were consistent for each period.

Since there are so many factors affecting seasonal variation which make it impossible to measure the variation with a high degree of accuracy, there is little justification for applying methods that are complex or that require lengthy computations.

---

[2] However, there were some slight changes in the seasonal variation from the 17-year period 1897–1913 to the 11-year period 1928–1938. During the earlier period, April was the highest spring peak; and during the later period, the peak came in both March and April. During both periods, September was the high month, but the level in the first period, 103 to 105, was somewhat lower than for the second, 107. These differences were due to changes in the industry, and not to methods.

TABLE 7.—COMPARISON OF SEASONAL INDEXES CALCULATED BY FOUR METHODS

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1928–1938

Month	Simple average	Moving average	Trend adjusted	Link relatives
January	95	92	95	93
February	99	99	100	97
March	103	103	103	102
April	103	103	103	102
May	101	103	101	101
June	100	101	100	100
July	102	102	102	102
August	104	105	104	106
September	107	107	107	107
October	100	101	100	103
November	95	95	94	97
December	91	90	91	90

Usually the choice of a method for calculating seasonal variation may rest with the ease of calculation. In the presence of secular trend, the simplest method which minimizes the effect can be used with a reasonable degree of accuracy.

USES

One of the important uses of seasonal indexes is the comparison of violence of and differences in seasonal movements. A student of agricultural economics should be familiar with the seasonal peculiarities of a wide range of commodities. Farming itself is a seasonal venture. The products are produced, stored, and sold seasonally; and the resulting prices also vary seasonally.

In the United States, wheat is harvested from May–June in Texas and Oklahoma to August–September in the Dakotas, whereas in the Southern Hemisphere wheat is harvested during our winter months.

The wheat inspection at Chicago gives some indication of the rates of marketing. Ten times as much wheat reached Chicago in August as in April. The bulk of wheat was marketed from July to October (table 8). Large quantities of wheat move to market because of the low moisture content which permits immediate shipment for consumption and storage at terminal markets and because of the weevil menace on farms.

The visible supply of wheat, that is the grain in public elevators, warehouses, in transit, etc., was generally lowest a few months after the months of low receipts, and high following the months of high receipts (table 8). However, the lowest visible supply occurred one month prior to the month with the highest current inspections. There was less violence in the seasonal changes in visible supply than in inspection. The index for visible supply ranged from 60 in July to 136 in January, whereas the inspection ranged from 28 in April to 278 in August (table 8).

TABLE 8.—SEASONAL VARIATION IN SUPPLIES, FUTURE TRADING, AND PRICES OF WHEAT

Month	Inspection	United States visible supply	World visible supply	Volume of future trading	Prices				
					Chicago	Mecklenburg	Flour, wholesale	Flour, retail	Bread, retail
January	48	136	118	74	99	101	97	97	100
February	31	133	115	65	99	101	102	102	99
March	35	128	115	77	99	101	101	102	99
April	28	120	110	110	100	99	102	101	99
May	34	104	97	96	104	98	106	103	99
June	29	84	87	120	102	100	99	100	99
July	174	60	78	148	96	100	100	99	100
August	278	61	74	129	96	99	102	102	101
September	203	62	80	93	100	99	96	100	101
October	155	87	97	105	101	99	96	97	101
November	108	105	111	102	101	101	97	97	101
December	76	120	119	80	103	102	102	100	101

The seasonal change in the world visible supply was much less than that for the United States. The United States visible was determined by the harvesting and marketing of our crop, whereas the world visible was determined by marketings influenced by a wide range of harvesting periods.

The volume of trading in wheat futures was highest just before the period of harvest marketing.

In spite of the violent seasonal variation in production, marketing, and storage, the prices were relatively non-variable. The index of the Chicago price was lowest in July and August, 96, and highest in May, 104. These variations were small compared with those for inspection and were in the opposite direction.

Prices of wheat in Mecklenburg, Germany, 125 years ago, also exhibited very little seasonal variation. Wholesale and retail prices of flour and retail prices of bread had little seasonal variation.

ELIMINATION OF SEASONAL VARIATION

Another important function of such indexes is in the elimination of seasonal variation from various types of data. A common method is to divide the data in their original form, or an index of it, by an index of seasonal variation. For example, the monthly prices of hogs for 1912 advanced from $6.30 in January to $8.65 in October, and then fell to $7.45 in December (table 9). Some of these monthly variations could not be ascribed to the peculiar characteristics of the year 1912. The normal seasonal variation for the year was eliminated by dividing the January price, $6.30, by the January seasonal index, 95; the February price, $6.25, by 99; and so on. The quotients, $6.63 for January, $6.31 for February, and so on, represent the prices adjusted for normal seasonal variation (table 9). The October to December decline was really not so drastic as the unadjusted prices indicated.

TABLE 9.—ELIMINATION OF SEASONAL VARIATION BY DIVISION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1912

Month	Price*	Seasonal variation†	Adjusted price
January	$6.30	95	$6.63
February	6.25	99	6.31
March	7.10	104	6.83
April	7.85	105	7.48
May	7.70	102	7.55
June	7.50	102	7.35
July	7.60	103	7.38
August	8.05	102	7.89
September	8.30	104	7.98
October	8.65	100	8.65
November	7.75	92	8.42
December	7.45	92	8.10

* Table 1, page 91. † Table 2, page 92, trend adjusted.

A still different method of eliminating seasonal variation is frequently used in the construction of index numbers. Many students of index numbers of farm prices eliminate seasonal variation by expressing the prices for a given month in terms of a base price for that month, rather than in terms of the average for the entire base period. The 1910–1914 United States farm price of butter averaged 25.6 cents per pound. The June and November 1939 prices were 23.8 and 27.3 cents, respectively. The index numbers for these two months in terms of the 5-year average, 25.6 cents, were 93 and 107, respectively. This would indicate that butter was very low in June compared with November. However, there was considerable difference in the June and November average prices during the base period. These averages were 23.5 and 26.7 cents, respectively. When the 1939 prices were compared with the corresponding monthly base prices, the index numbers for June and November were approximately the same,[3] 101 and 102. These adjusted indexes were a much better basis of comparison of June and November prices than the unadjusted indexes, 93 and 107. This is a rather simple and usually a very effective way of eliminating seasonal variation from index numbers.

The normal seasonal variation in some products, such as milk and eggs, has changed decidedly with passing time. In spite of the use of the above method, the indexes would be incorrect if there were a change in the seasonal variation between the base period and the present. The 1910–1914 New York farm price of milk averaged $1.91 per 100 lb. for January; and $1.05 for June. The corresponding averages for 1933–1937 were $1.74 and $1.50, respectively. Obviously, there was a decided change in the relative prices for the two months. The elimination of seasonal variation in any one year would result in different relative prices for January and June, depending on which of the periods was used as a base for seasonal variation. This type of change may be sudden, but is usually gradual and unnoticed for a considerable time after it has begun.

Sometimes, the change is mostly in the violence of variations; at other times there is a shift in the position of the highest or lowest months. When either of these changes occurs, the methods here described are not applicable. Students have devised variations of these methods to fit the problem at hand.[4]

Another factor affecting the usefulness of seasonal indexes is the degree of irregularity in the index. The seasonal index is an average of a number of years in which seasonal variation may be markedly different because of size of crops, strikes, weather, and other acts of man and Providence. When the index is so influenced, the seasonal movements of one period may not be present in other periods or even in every year of the same period.[5] The reliability of the index for any purpose depends on homogeneity within months. All too often this problem is overlooked.

---

[3] The June 1939 price, 23.8 cents, divided by the June 1910–1914 price, 23.5 cents, equals 1.01. The November 1939 and November base prices were both higher than June prices, but their ratio was approximately the same as for June (27.3 ÷ 26.7 = 1.02).

[4] Spencer, L., and Pearson, F. A., A New Index of Milk Prices in New York, Farm Economics No. 86, pp. 2089–93, June 1934.

Spencer, L., A Revised Series of Milk Prices for New York, Farm Economics No. 111, pp. 2707–10, February 1939.

---

[5] Waite, W. C., Cox, R. W., Seasonal Variations of Prices and Marketings of Minnesota Agricultural Products, 1921–1935, University of Minnesota Agricultural Experiment Station, Technical Bulletin 127, March 1938.

Thomsen, F. L., Agricultural Prices, pp. 259–260, 1936.

CHAPTER 7

CYCLES

Much statistical work concerns the removal of secular trend and seasonal variation from time series in order to study cycles, the influence of supply on price, and other problems. Most treatments of this subject emphasize the elimination of trend and seasonal variation from monthly data. Most of the problems of the agricultural economist, however, center about annual, rather than monthly series. For this reason, this chapter will emphasize the removal of trend from annual series.

TABLE 1.—FIRST-DIFFERENCE, PERCENTAGE-OF-PRECEDING-YEAR AND LEAST-SQUARES METHODS OF ELIMINATING TREND

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Year	Price of hogs*	First-difference method	Percentage-of-preceding-year method	Least-squares method	
				Trend, $Y = 6.034 + 0.248x$	Price in per cent of trend
1896	$3.39	—	—	—	—
1897	3.64	$ +0.25	107	$4.05	90
1898	3.85	+0.21	106	4.30	90
1899	4.03	+0.18	105	4.55	89
1900	5.05	+1.02	125	4.79	105
1901	5.89	+0.84	117	5.04	117
1902	6.93	+1.04	118	5.29	131
1903	6.00	−0.93	87	5.54	108
1904	5.15	−0.85	86	5.79	89
1905	5.22	+0.07	101	6.03	87
1906	6.25	+1.03	120	6.28	100
1907	6.04	−0.21	97	6.53	92
1908	5.74	−0.30	95	6.78	85
1909	7.50	+1.76	131	7.03	107
1910	8.88	+1.38	118	7.27	122
1911	6.63	−2.25	75	7.52	88
1912	7.54	+0.91	114	7.77	97
1913	8.23	+0.69	109	8.02	103

* Calculated from table 1, page 91.

ANNUAL SERIES

First Differences

First differences are the simplest method of removing secular trend from annual series. Their calculation is easy, and the method is effective. If cycles or other fluctuations in which the student is interested are present, they are likely to be detected.

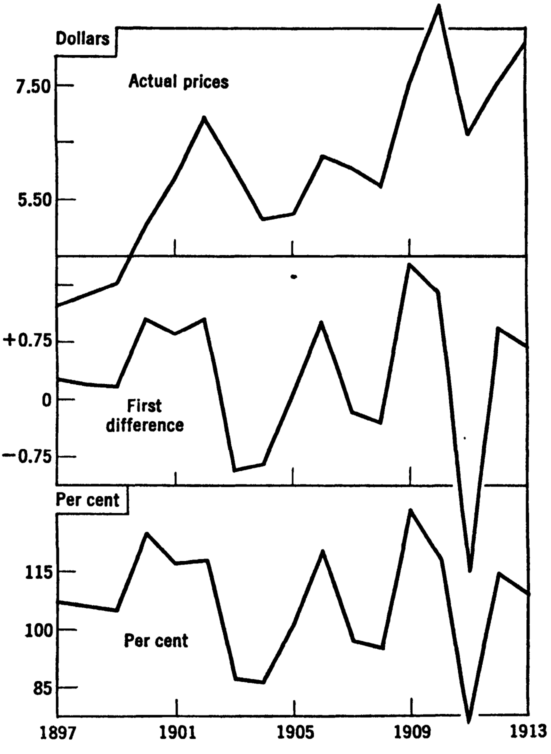


FIGURE 1.—CYCLES SHOWN BY FIRST-DIFFERENCE AND PERCENTAGE-OF-PRECEDING-YEAR METHODS OF ELIMINATING TREND[1]

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Both methods eliminate most of the trend. The lines indicate the presence of similar cyclical fluctuations.

The prices of hogs in 1896 and 1897 were $3.39 and $3.64, respectively. The first difference for the year 1897 was + $0.25 (3.64 − 3.39 = 0.25). The first difference for 1898 was + 0.21; for 1899, + 0.18, and so on

---

[1] From table 1, page 104.

(table 1). These differences indicate that prices were increasing and relatively high during the periods centering around 1901, 1906, and 1910. They were decreasing and low about 1903–1904, 1907–1908, and 1911. When these first differences were plotted, it was clear that most of the trend was eliminated (figure 1).

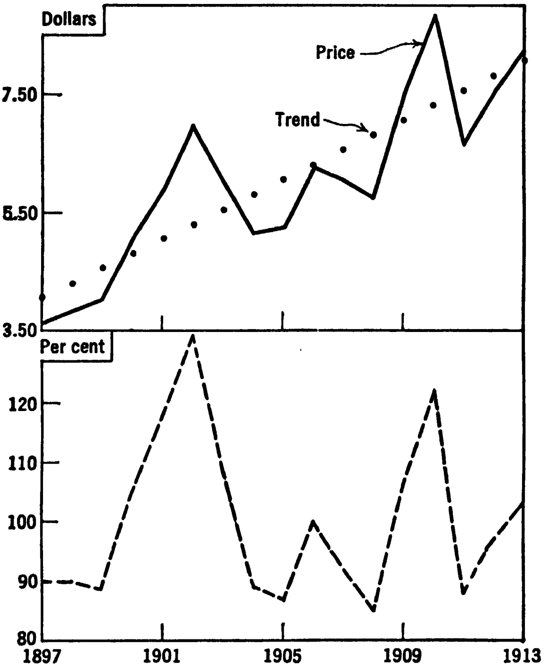


FIGURE 2.—CYCLES SHOWN BY LEAST-SQUARES METHOD OF MEASURING TREND[2]

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

The dots lying in a straight line represent the linear trend of the price of hogs, the solid line.

The broken line below represents the percentage the price of hogs was of the trend. The fluctuations were cyclical.

PERCENTAGE OF PRECEDING YEAR

Another simple method of eliminating trend is to express each year as a percentage of the preceding year. The 1897 price of hogs, $3.64, was 107 per cent of that for the preceding year (3.64 ÷ 3.39 = 1.07). The percentages for 1898 and 1899 were 106 and 105 (table 1). When the percentages were more than 100, the hogs were rising in price; and when less, they were falling. This method eliminated substantially all the trend (figure 1).

---

[2] Table 1, page 104.

The principle of percentage of preceding year is very similar to that of first differences. First differences are absolute differences and are expressed as dollars, cents, tons, or other units, positive or negative. Percentages of preceding year show the relative differences, and declines and advances are indicated by the relation of the percentage to 100. The results of the two procedures are quite similar when plotted on their respective scales (figure 1). Both methods show change rather than absolute values.

TABLE 2.—MOVING-AVERAGE METHOD OF ELIMINATING TREND

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Year	Price of hogs*	Moving averages			Per cent of moving average		
		3-year	5-year	7-year	3-year	5-year	7-year
1897	$3.64	$3.63†	$3.85†	$4.19†	100	95	87
1898	3.85	3.84	3.99†	4.31†	100	96	89
1899	4.03	4.31	4.49	4.68†	94	90	86
1900	5.05	4.99	5.15	5.06	101	98	100
1901	5.89	5.96	5.58	5.27	99	106	112
1902	6.93	6.27	5.80	5.47	111	119	127
1903	6.00	6.03	5.84	5.78	100	103	104
1904	5.15	5.46	5.91	5.93	94	87	87
1905	5.22	5.54	5.73	5.90	94	91	88
1906	6.25	5.84	5.68	5.99	107	110	104
1907	6.04	6.01	6.15	6.40	100	98	94
1908	5.74	6.43	6.88	6.61	89	83	87
1909	7.50	7.37	6.96	6.94	102	108	108
1910	8.88	7.67	7.26	7.22	116	122	123
1911	6.63	7.68	7.76	7.53†	86	85	88
1912	7.54	7.47	7.90†	7.71†	101	95	98
1913	8.23	7.99†	7.52†	8.01†	103	109	103

* Table 1, page 91.   † Based on prices prior to 1897, or following 1913.

PERCENTAGE OF STRAIGHT-LINE TREND

A commonly used method of eliminating trend expresses each item in the series as a percentage of the straight-line trend. The straight-line trend may be estimated or fitted by a variety of methods as outlined in chapter 5. From the straight line, the trend values for each year may be determined. From the least-squares line, the estimated price of hogs for 1897 was $4.05 (table 1). The ratio of the actual price, $3.64, to the trend price was 90 per cent (3.64 ÷ 4.05 = 0.90). The price rose to $3.85 in 1898, but the trend was also upward, and the ratio was again

90. By 1913, the price had risen to $8.23, and the trend to $8.02, and the ratio was 103 (table 1). These percentages indicate that, after the elimination of trend, the price of hogs was highest in 1901–1902, 1906, and 1910; and lowest in 1899, 1904–1905, 1908, and 1911 (figure 2). These percentages are in terms of values, not rates of change as were the first differences. The method is well adapted to elimination of linear trend and reveals the presence of definite cycles in hog prices (figure 2).

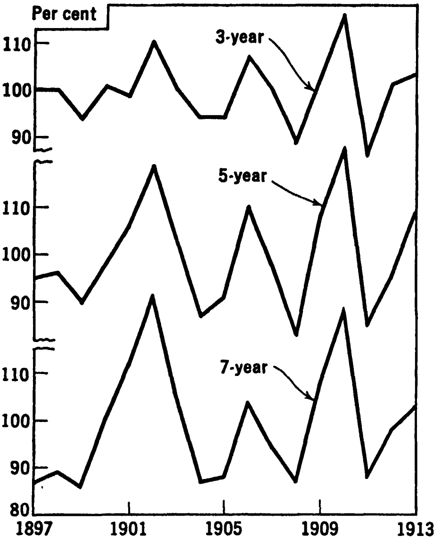


FIGURE 3.—CYCLES SHOWN BY MOVING-AVERAGE METHOD OF ELIMINATING TREND

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

The cycles are most clearly shown by the use of the 7-year moving average. The 3-year moving average removes not only trend but also a considerable part of the cycle. The 5-year average removes all the trend and a small part of the cycle.

PERCENTAGE OF MOVING AVERAGES

As was stated in chapter 5, moving averages may be used to eliminate trend. Each price may be expressed as a percentage of the moving average. The 1902 price of hogs, $6.93, was 111 per cent of the 3-year moving average centered on 1902 (6.93 ÷ 6.27 = 1.11). The percentages of the 5- and 7-year centered moving averages were 119 and 127 (table 2). The fluctuations of percentages of 3-year averages were less violent than those based on 5- or 7-year averages (figure 3). The 3-year averages were more like the prices for the years on which they were centered than were 5- and 7-year averages. This is clearly shown when the three moving averages are plotted with the original prices (figure 6, chapter 5, page 86).

All the moving averages eliminated secular trend (figure 3); the 3-year moving average, in addition, removed some of the cyclical variation. The process of dividing a series containing a cycle by another series also containing most of the same cycle will result in a series which will contain very little of it. For the period covered in this example, even a 5-year average was a little too short to approximate a smooth curve, and its use also eliminated a little of the cyclical variation. The 7-year moving average was long enough so that it was not influenced by cyclical variations. The division of the hog prices, which contained cycles, by the 7-year average, in which the cycles have been ironed out, gave percentages which contained all the original cyclical variations.

A comparison of the percentages of the different moving averages reveals that the 7-year-average method shows all the elements of the cycle and resembles the percentages of a straight line (figures 2 and 3). When the trends are not linear, it is more satisfactory to express prices as a percentage of a moving average than of straight lines. However, the student must use care in the selection of the moving average.

PURCHASING-POWER METHOD

A very common method of eliminating trend from the prices of a commodity is to divide the series by the prices of one or more other commodities. If the prices are divided by an index of some group of commodities, the result is called "purchasing power" by some students and "deflated series" by others.[3]

Prices of individual commodities are subject to monetary forces common to all. In an index of many commodities, most fluctuations peculiar to individual commodities tend to be ironed out. Such an index shows only the result of common forces.

In periods of rapid change, some prices change more rapidly and to a greater extent than others. Therefore, the divisor index should be about as flexible as the price of the product to be adjusted. Consequently, the choice of the divisor depends upon the particular price series. For example, index numbers of farm prices of all farm products or wholesale prices of 30 basic commodities are better divisors for highly flexible farm and wholesale prices of live hogs than are indexes of retail prices of foods or of wholesale prices of all commodities.

All the trend is eliminated from a purchasing-power series when the price to be analyzed and the divisor price have approximately the same trend.

The price of hogs for the year 1897, $3.64, was divided by the United States Bureau of Labor index of wholesale prices of farm products, 60

---

[3] The term "deflated series," strictly construed, means that the level of the individual series has been reduced. This expression was introduced during the period of rising or high prices, 1914–1929. In a period of generally declining prices, such as 1865–1896, the level of the individual series would very likely be raised. Strictly speaking, this would not be a deflated series. It would be an inflated series. The expression "purchasing power" is a better term because, unlike "deflated series," it does not imply that the prices will be lowered. An additional advantage of the term "purchasing power" is that it denotes the exact or relative amount of one or more commodities that a given commodity or commodities will buy.

(table 3). The quotient, $6.07, represents the purchasing power in dollars. The purchasing power in dollars for 1898 was $6.11 (3.85 ÷ 63 = 0.0611). The 1902 purchasing power, $8.45, was quite similar to those for 1910 and 1913, $8.54 and $8.23.

TABLE 3.—PURCHASING-POWER METHOD OF ELIMINATING TREND

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

Year	Price of hogs*	Purchasing power in terms of many commodities		Purchasing power in terms of one commodity	
		Index of wholesale prices of all farm products, 1910–1914 = 100	Purchasing power	Price of corn per bushel	Purchasing power, hog-corn ratio
1897	$3.64	60	$6.07	$0.26	14.0
1898	3.85	63	6.11	0.32	12.0
1899	4.03	64	6.30	0.33	12.2
1900	5.05	71	7.11	0.38	13.3
1901	5.89	74	7.96	0.50	11.8
1902	6.93	82	8.45	0.60	11.6
1903	6.00	78	7.69	0.46	13.0
1904	5.15	82	6.28	0.51	10.1
1905	5.22	79	6.61	0.50	10.4
1906	6.25	80	7.81	0.46	13.6
1907	6.04	87	6.94	0.53	11.4
1908	5.74	87	6.60	0.69	8.3
1909	7.50	98	7.65	0.67	11.2
1910	8.88	104	8.54	0.58	15.3
1911	6.63	94	7.05	0.59	11.2
1912	7.54	102	7.39	0.69	10.9
1913	8.23	100	8.23	0.63	13.1

* Table 1, page 104.

The purchasing power of hogs reveals about the same cyclical fluctuations shown by other methods of eliminating trend (compare figure 4 with 2 and 3). Not quite all the trend was eliminated by the purchasing-power method because prices of farm products advanced somewhat less rapidly than the price of hogs.

The price of a commodity may also be divided by and expressed as a ratio to another individual commodity. The hog-corn ratio is a case in point. In 1897, the Chicago price of hogs was $3.64 per 100 pounds, and the Chicago price of corn $0.26 per bushel (table 3). Therefore, 100 pounds of hogs were worth 14.0 bushels of corn (3.64 ÷ 0.26 = 14.0). The hog-corn ratio, 14.0, was the number of bushels of corn equal to 100 pounds of hogs in 1897. The method eliminated most of the trend (figure 4). From 1904 to 1913, the cycles were about the same as those obtained by other methods (compare figure 4 with 1, 2, and 3). Prior to 1901, the cycles were disturbed by the abnormally low prices for corn.

This illustrates an important principle involving the use of an individual commodity as a divisor. Although it may eliminate most of the trend, it may also inject its peculiar fluctuations into the series studied. In extreme cases, the fluctuations of the divisor may entirely obscure the characteristics of the original series.

For some types of study, the hog-corn ratio is very valuable, but it is not very satisfactory for a cyclical analysis of hog prices.

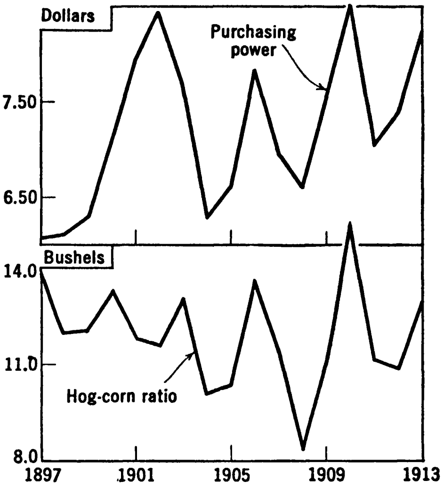


FIGURE 4.—CYCLES SHOWN BY THE PURCHASING POWER OF HOG PRICES IN TERMS OF AN INDEX OF ALL FARM PRODUCTS, AND IN TERMS OF THE PRICE OF CORN, 1897–1913

The purchasing power of hogs in terms of all farm products (above) did not eliminate all the trend. The purchasing power in terms of corn, or the hog-corn ratio (below), eliminated all the trend but obscured the cyclical fluctuations centering about 1902.

USES

In the analysis of many price problems, any one or several of the above methods may be employed. The relation of the number of cattle to their price is best examined when three different methods of analysis are used.

The prices of cattle from 1880 to 1937 were expressed as a purchasing power. The fluctuations due to movements in the general price level were removed by dividing by an index of prices of all commodities. The results indicated that most of the trend had been eliminated and violent fluctuations in original prices had been reduced to a very regular cycle of high and low prices (figure 5). The actual prices indicated that cycles might be present but gave an incorrect impression of their magnitude and length. The peaks in prices were not of the same height or the same distance apart as the peaks in purchasing power (figure 5).

For the first part of the period, expressing prices as percentages of trend or first differences would have removed the trend satisfactorily. However, in the last part when the price level fluctuated violently after 1914, these methods would not have been satisfactory. The purchasing-power method proved satisfactory for the entire period.

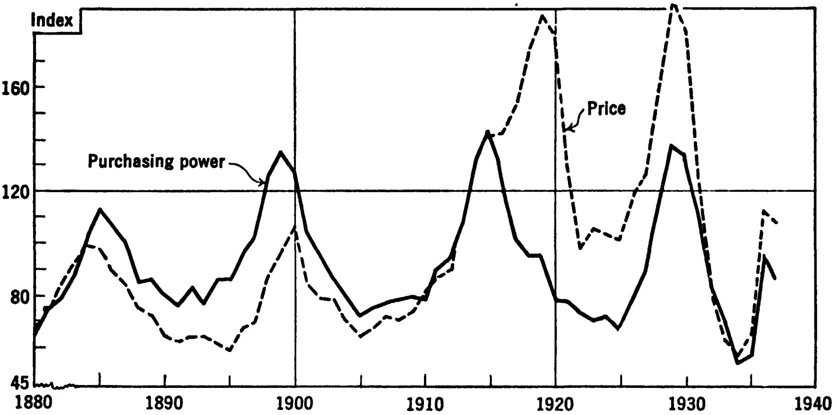


FIGURE 5.—CYCLES SHOWN BY INDEX NUMBERS OF ACTUAL PRICES AND PURCHASING POWER

UNITED STATES FARM PRICE OF BEEF CATTLE, JANUARY 1, 1880–1937
1910–1914 = 100

The purchasing power shows a very clear, regular cycle, which is somewhat obscured in the actual prices. This is particularly true during the violent gyrations from 1914 to 1937.

Throughout this period, there was a general increase in the number of cattle on farms. The trend was eliminated by expressing the number of cattle as a percentage of the least-squares trend line. The resulting series showed very regular cycles in cattle numbers (figure 6). This series appears to be related to the purchasing power of cattle prices (figure 6). In general, the purchasing power was high when numbers were low. However, the peaks in purchasing power occurred some time after the low points in total numbers had been reached. Conversely, the low points in purchasing power came when numbers were declining.

The application of the first-difference method to the numbers of cattle brings out this relationship more vividly (figure 7). The annual first differences represent the change that occurred in the number of cattle. The greatest positive first difference occurred in the year when

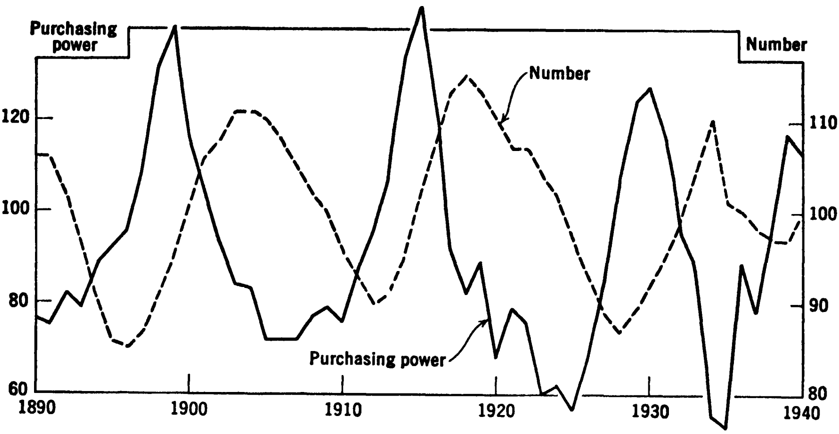


FIGURE 6.—CYCLES IN THE NUMBER OF CATTLE EXPRESSED IN PERCENTAGE OF TREND AND THE PURCHASING POWER OF THE PRICES, JANUARY 1, 1890–1940

Both methods disclose fairly regular cycles. When the number was large, prices tended to be low. However, the one curve was not the exact reverse of the other.

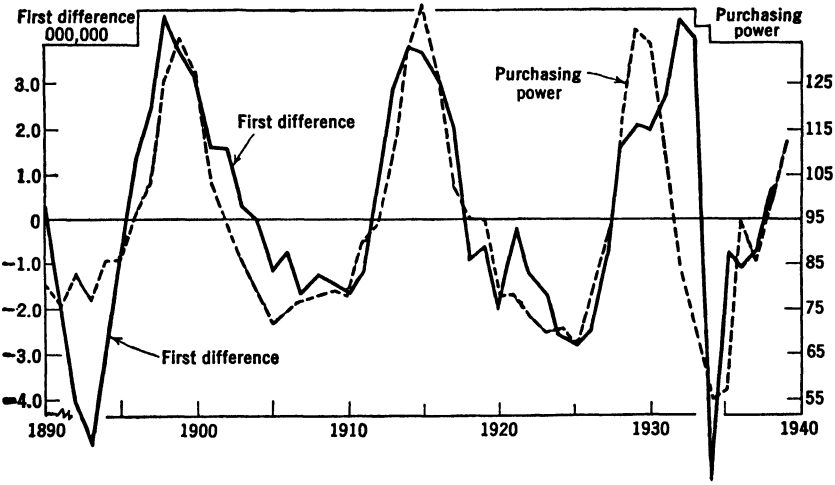


FIGURE 7.—CYCLES SHOWN BY THE FIRST DIFFERENCES OF THE NUMBER OF CATTLE AND THE PURCHASING POWER OF THE PRICES, JANUARY 1, 1890–1939

These methods reveal very regular cycles in the purchasing power and in the first difference—the annual change in the number of cattle on farms. The change in the number is more closely related to price than the actual number.[4]

---

[4] Compare with figure 6.

numbers increased the most; and the greatest negative difference, when numbers decreased the most. When the change in numbers, as measured by first differences, and the price, expressed as a purchasing power, are plotted on the same graph, their close relationship is very evident (figure 7). This fact offers a clue to the nature of the relationship between numbers and prices of cattle. Prices appear to be affected more by breeding demands, as reflected by increase in numbers, than by the actual number on hand. The peak in purchasing power occurred when the number was increasing most rapidly, rather than when the number was at its lowest point. This relationship between change in numbers and purchasing power could be ascertained from figure 6, which shows the actual numbers and purchasing power, provided that intensive study of the graph were made. The use of first differences brings out this relationship in bold relief.

In an analysis like the above, the student should be fully equipped with all possible methods of attack and should not hesitate to use any or all of them. The particular method which suits the problem cannot always be detected prior to the attack, and some experimenting is required.

A very usual type of problem in eliminating trend arises in the analysis of the relationship between production and price of farm crops. The analysis is especially difficult for the periods of widely gyrating prices and shifting production experienced since 1914. Expressing prices as purchasing power is about the only satisfactory method of eliminating wild fluctuations due to the price level. For some products, even this procedure does not eliminate all the trend because there is trend in the purchasing power itself. When this is true, the purchasing power may be expressed as a percentage of a moving average or some other type of trend.

The problem of eliminating trend from production data is almost as difficult. Straight lines rarely describe production during this period satisfactorily. To make the production series comparable to the purchasing-power series, production is frequently expressed as a percentage of a moving average. It is also possible to eliminate trends in the comparison of production and prices by calculating their first differences or expressing each as a percentage of the preceding year. However, year-to-year relationships are less obvious when series are adjusted by this method than when the trend is eliminated by moving averages. The variability in first differences is relatively greater than variability in absolute values.

Problems of this nature emphasize further the necessity of having a wide knowledge of a number of different methods.

COMPARISON

In general, first differences are the easiest to calculate of all trend-adjusted series. Moreover, first differences usually show the presence of cycles or shorter time variations. Sometimes first differences are even superior to other methods.

Percentages of preceding years are very similar in principle to first differences but somewhat more difficult to calculate. Expressing series as a percentage of trend is often a satisfactory method of obtaining a series comparable to the original but with trend eliminated. Whether a straight line, curves, or moving averages are employed depends upon the nature of the trend. Moving averages are the most flexible and are widely used.

In short periods of general price stability, the trend in prices of individual series may be eliminated by one of the above methods. However, since 1914, prices have fluctuated so violently over such short periods that these methods do not satisfactorily eliminate variations due to the price level. It is necessary to adjust these data by dividing by a general index of prices. When there is a trend in purchasing power, further adjustment is sometimes desirable.

MONTHLY SERIES

To the agricultural statistician, the analysis of monthly series is not so important a problem as the analysis of annual series. Nevertheless, he occasionally examines monthly data for cycles, supply-price relationships, and the influences of other factors. In such analyses, the problem of eliminating trends is much the same with monthly as with annual data. In addition, there is often the further problem of eliminating seasonal variation. The methods of analyzing monthly data are merely combinations of previous methods studied for the elimination of trend and seasonal variation. Some methods eliminate trend first; some, seasonal variation first; and others, both simultaneously.

PERCENTAGE OF CORRESPONDING MONTH OF PREVIOUS YEAR

A method of eliminating trend and seasonal variation in one operation is the division of monthly data by the corresponding monthly values for the preceding year.

The price of hogs in January 1911 was $7.85, or 90 per cent of the January 1910 price, $8.70 (table 4). Similarly, the February 1911 price, $7.25, was 79 per cent of the February 1910 price, $9.20. Although these percentages are measures of change rather than the level of hog prices, they indicate the presence of cycles (figure 8). Dividing the

TABLE 4.—PERCENTAGE-OF-CORRESPONDING-MONTH-OF-PRECEDING-YEAR METHOD OF ELIMINATING TREND AND SEASON IN ONE OPERATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1911–1912

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Price, dollars												
1910....	8.70	9.20	10.65	10.00	9.50	9.35	8.60	8.25	8.70	8.45	7.55	7.65
1911....	7.85	7.25	6.70	6.15	5.85	6.15	6.65	7.15	6.75	6.50	6.35	6.25
1912....	6.30	6.25	7.10	7.85	7.70	7.50	7.60	8.05	8.30	8.65	7.75	7.45
Per cent												
1911....	90	79	63	62	62	66	77	87	78	77	84	82
1912....	80	86	106	128	132	122	114	113	123	133	122	119

prices by those of the previous year removes most of the trend. Dividing the price for each month by that for the corresponding month would have eliminated all seasonal variations if there had been no variation from year to year in the amount or time of the seasonal movements. However, the influence of seasonal variation was not constant from year to year, and the resulting percentages show some monthly fluctuations.

TABLE 5.—MOVING-AVERAGE AND NORMAL-SEASONAL METHOD OF ELIMINATING TREND AND SEASON IN TWO OPERATIONS

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1910

Month	Price of hogs*	Moving average†	Percentage of moving average	Normal season‡	Cycle
January.............	$ 8.70	7.09	123	95	129
February............	9.20	7.12	129	99	130
March..............	10.65	7.14	149	104	143
April................	10.00	7.17	139	105	132
May................	9.50	7.19	132	102	129
June................	9.35	7.21	130	102	127
July................	8.60	7.22	119	103	116
August..............	8.25	7.24	114	102	112
September...........	8.70	7.26	120	104	115
October.............	8.45	7.29	116	100	116
November...........	7.55	7.31	103	92	112
December...........	7.65	7.33	104	92	113

* Table 1, page 91. † 84-month moving average centered on 43rd month. ‡ Page 92.

MOVING-AVERAGE METHOD

Annual data were adjusted for trend by expressing them as a percentage of the moving average. The same method may be applied to monthly data with a device for eliminating seasonal variation. The 7-year or 84-month moving average was calculated for the price of hogs. The January 1910 price of hogs, $8.70, was 123 per cent of the centered moving average, $7.09. Likewise, the February price was 129 per cent of its moving average (table 5). Although the trend had been removed from these percentages, any seasonal variation in the original series was still present. This was removed by dividing each percentage by an index of seasonal variation. The January 1910 per cent, 123, was divided by the January index of seasonal variation, 95. The quotient, 129, represents the price with both trend and seasonal variation removed (table 5). The same procedure was followed throughout each year of the series. This method results in clear cycles in hog prices (figure 8). They are shown as the level of hog prices rather than the change in price. Some seasonal variation still remains because the normal seasonal variation does not prevail throughout every season. However, the normal seasonal variation is more likely to fit a particular year than the seasonal variation of the preceding year. For this reason, the moving-average method removes seasonal variation more accurately than the percentage-of-the-corresponding-month-of-previous-year method.

Several variations of this general method have been employed. Straight lines and other rigid curves have been used in place of moving averages.

Instead of dividing the percentages of trend by the index of normal seasonal variation, the seasonal index is often subtracted from the percentages.

PURCHASING-POWER METHOD

As previously explained, price series contain fluctuations due to the general movement of all prices, which are difficult to remove with moving averages, straight-line trends, first differences, and the like. The influence of the price level and of seasonal variation on monthly data may be removed by calculating an index of the prices in terms of corresponding months of the base period and deflating this seasonally adjusted index. Prices of hogs in 1910–1911 were expressed as a percentage of the average prices for the corresponding months of the 5 years 1910–1914. The January 1910 price, $8.70, was 113 per cent of the January 1910–1914 price, $7.72. Likewise, the adjusted index for February 1910 was 117 (table 6). The other months were treated in the same manner. The resulting index numbers were adjusted for season, but not for changes in the price level. Adjustment for changes in the price level was made by dividing the seasonally adjusted index for each month by an index of wholesale prices of 30 basic commodities. The index for January 1910 divided by the index of 30 basic commodities, 106, gave the purchasing power of hogs, 107. For February 1910, the index of prices of 30 basic

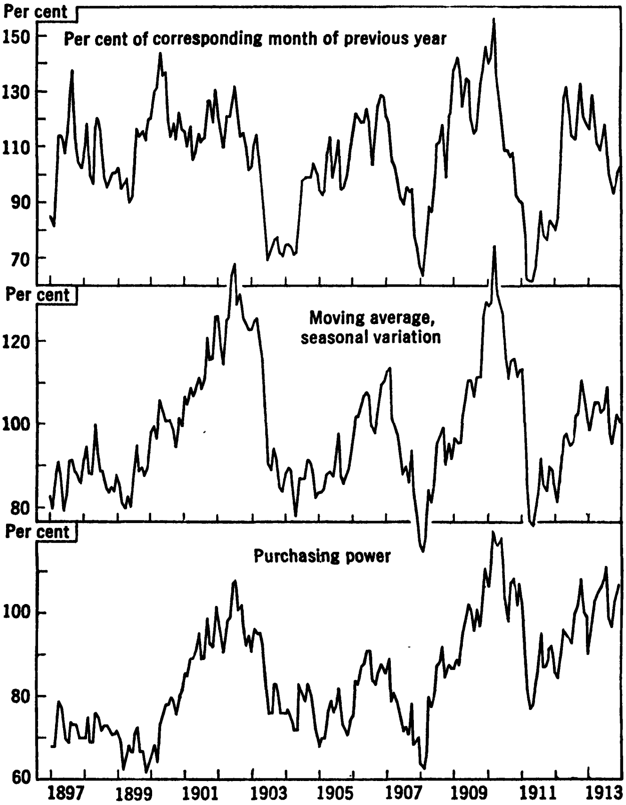


FIGURE 8.—CYCLES SHOWN IN THE MONTHLY PRICES OF HOGS BY THREE METHODS OF ELIMINATING TREND AND SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO, 1897–1913

The three methods show about the same cyclical fluctuations in the monthly prices of hogs. The purchasing power method did not eliminate so much of the trend as the other methods.

commodities was 104; and the purchasing power of hogs, 112. The plotted index of purchasing power disclosed a distinct cycle in hog prices. The trend was not entirely removed by the divisor (figure 8). Except for the small amount of trend remaining, the cycle is quite similar to that obtained with the moving-average method.

There are variations in the purchasing-power method. Seasonal

TABLE 6.—PURCHASING-POWER METHOD OF ELIMINATING TREND AND SEASONAL VARIATION

WHOLESALE PRICES OF HEAVY HOGS AT CHICAGO IN TERMS OF 30 BASIC COMMODITIES, 1910–1911

Months	1910–1914 base price of hogs*	1910				1911			
		Price of hogs*	Index, 1910–1914 = 100	Divisor, index of 30 basic commodities	Purchasing power	Price of hogs*	Index, 1910–1914 = 100	Divisor, index of 30 basic commodities	Purchasing power
January.....	$7.72	$ 8.70	113	106	107	$7.85	102	99	103
February ...	7.86	9.20	117	104	113	7.25	92	98	94
March......	8.36	10.65	127	106	120	6.70	80	97	82
April.......	8.26	10.00	121	104	116	6.15	74	96	77
May.......	7.95	9.50	119	102	117	5.85	74	95	78
June.......	7.93	9.35	118	100	118	6.15	78	94	83
July........	8.08	8.60	106	102	104	6.65	82	94	87
August.....	8.06	8.25	102	103	99	7.15	89	94	95
September..	8.09	8.70	108	101	107	6.75	83	95	87
October.....	7.86	8.45	108	100	108	6.50	83	95	87
November ..	7.39	7.55	102	99	103	6.35	86	94	91
December...	7.23	7.65	106	99	107	6.25	86	93	92

* Table 1, page 91.

variation may be eliminated after the purchasing power for each month has been calculated. The seasonal adjustment may consist of expressing the purchasing power as a percentage of the corresponding month of the same base period, or dividing it by an index of seasonal variation.

USES

The purposes of analyzing monthly data are the same as for annual data. Monthly data are used to study the influence of certain factors over shorter periods than one year and to gauge more closely the timing of certain changes. Since agricultural production is mostly on an annual basis and prices are strongly governed by annual factors, the analysis of monthly data is somewhat restricted. A relatively more important field of monthly analysis is in business activity and the like. Index numbers of business activity are published with and without seasonal adjustment, and with and without the elimination of trend.

CHAPTER 8

TABULAR ANALYSIS OF RELATIONSHIPS

In the study of individual variables, attention was directed to central tendency and dispersion. The problem of analyzing relationships is nothing more than comparing the variations in one series with those in another series. For example, it is known that farms vary in size, and it is known that incomes vary. The question arises whether the variations in the size of farms are related to the variations in income. Are incomes any more or any less on large farms than on small farms? This is a simple statement of the problem of relationships.

The problem of relationships is very real and very commonplace. Every individual from daylight to dawn observes relationships, accepts relationships, and makes decisions on the basis of relationships. Because relationships affect every human endeavor, they are the most important problem of statistics. In fact, they are the main reason for the study of statistics.

TABULAR METHOD

Tabulation is one of the most elementary methods of analyzing relationships. It is so elementary that most textbooks either ignore the subject or bury it in a mass of detail concerning the construction of tables. This lack of emphasis on tabulation is due to its seeming simplicity rather than to its lack of importance. In fact, the tabular method is the most important and widely used method of analyzing relationships.

In a simple form, the tabular method involves classification of the data into groups according to one factor and the calculation of averages of a second factor for these groups. The procedure involves a few relatively simple steps. These steps can be easily followed in one-way tabular analysis.

ONE-WAY TABULAR ANALYSIS

1. The observations are divided into several groups on the basis of the size of one of the factors. In the study of frequency distributions,[1]

---

[1] Page 2.

these groups are called "classes." Each class or group includes all the observations where the factor in question falls within a certain range, for example "8 to 10" years of age, "20 to 29" cows per farm.[2]

TABLE 1.—ONE-WAY TABULATION

RELATION OF CROP YIELDS TO INCOME, 907 NEW YORK FARMS, 1927

Class intervals, index of yields	Frequency, number of farms	Totals for second factor, incomes	Average income
Less than 60	78	$ 11,308	$ 145
60– 79	153	26,222	171
80– 99	263	66,091	251
100–119	259	103,779	401
120–139	109	87,901	806
140 or more	45	45,197	1,004

To illustrate one-way tabular analysis, the relation of crop yields to incomes on 907 farms was studied. The indexes of yields were grouped into 6 classes (table 1). The first class included all farms with an index of yields less than 60. The highest and lowest classes were open at the ends. The range within each of the 4 intermediate classes was 20.

2. The total number of observations in each class is obtained. The number of farms in the first class was 78; that is, there were 78 farms with a crop index of less than 60 (table 1).

3. The total of all the values for a second factor is obtained for each class. The total of the 78 incomes for the first class was $11,308.

TABLE 2.—ONE-WAY TABLE

RELATION OF CROP YIELDS TO INCOME, 907 NEW YORK FARMS, 1927

Yields, index	Income
Less than 60	$ 145
60– 79	171
80– 99	251
100–119	401
120–139	806
140 or more	1,004

4. For each class, the total of the second factor is divided by the number of observations in the class. The results are a series of simple

---

[2] In making class intervals for frequency distributions, there were rather rigid rules concerning the number of classes, class limits, indeterminate classes, and the like. In tabular analysis, most of these principles apply, but the number of classes is usually smaller, and there is much greater use of unequal classes and indeterminate, or "open-end," classes.

averages of the second factor. The total income for the first class, \$11,308, was divided by the number of farms, 78. The average income for these farms with poor crop yields was \$145.

5. The classes, which are based on the first factor, and the averages, which are based on the second factor, are then arranged in a simple table. The classes of crop yields were arranged from poorest to best, and the corresponding averages for income were set down in an adjoining column (table 2).

6. The relationship is examined by comparing the averages[3] for the second factor for different values of the first factor. A simple one-way table clearly answers two questions in which the research worker is interested: (1) whether any relationship exists, and if so (2) whether the relationship is positive or negative—that is, whether the second factor increases or decreases as the first increases. A relationship exists when the averages show either a consistent increase or a consistent decrease. Whether the relation is positive or negative can be determined by reading down the column and noting whether the averages increase or decrease.

Comparison of average incomes for farms with different yields indicated the existence of a relationship between crop yields and income. The relationship was positive; that is, incomes rose as crop yields became better.

INDEPENDENT AND DEPENDENT VARIABLES

Ordinarily, research workers discuss relationships between yields and income, age and death rate, and the like. The statistician tends to generalize and speak in terms of independent and dependent variables, often using terms such as X_1 and X_2 to designate them. In the relationship shown in table 2, income may be called X_1, or the dependent variable. The term "dependent" merely connotes that the statistician assumes that income is dependent on variations in yield. Similarly, the statistician calls crop yields the independent variable, or X_2.

In the tabular method, the usual procedure is to group or classify the observations according to the independent variable and to add and average the numerical values for the dependent variable. This is not a rigid rule. Some workers classify the dependent and average the independent variable. Relations can be observed by either procedure.

Frequently, data are grouped by an independent variable, and *several* dependent variables are averaged.

---

[3] Though the most usual description of the second factor is the arithmetic mean, other measures, such as sums, frequencies, percentages, ratios, medians, standard deviations, and coefficients of variability, are sometimes used.

TWO-WAY TABULAR ANALYSIS

A two-way table shows the relationship of two independent variables to one dependent variable. The variation in the dependent variable is said to depend on the two independent variables.

In making a two-way table, the data are first classified according to one of the independent variables. Then, each group is subdivided according to the second independent variable. The values of the dependent variable are then averaged for all the combinations of the two independent variables.

TABLE 3.—TWO-WAY TABLE

RELATION OF CROP YIELDS AND SIZE OF FARM TO INCOME, 907 NEW YORK FARMS, 1927

Yields, index	Size of farm, units		
	Less than 215	215 to 344	345 or more
	Income	*Income*	*Income*
Less than 90	\$−110	\$+238	\$+ 600
90–109	+ 31	+158	+ 711
110 or more	+202	+500	+1,261

The 907 farms were first classified by crop yields into three approximately equal groups.[4] Each of these groups was then subdivided into three subgroups on the basis of the second independent variable, size of farm. The class limits for size of farm are given across the top of the table. At this point there were nine groups, the largest of which contained 139 farms, and the smallest, 84. The incomes for each group were added and averaged. An orderly arrangement of these averages is shown in table 3.

Such a table enables one to analyze the relationships of crop yields and the size of farm to income. There are several ways to look at such a table: the columns can be read vertically, the rows horizontally, or the whole table diagonally. When the table is read down, it appears that incomes generally increase with crop yields, regardless of the size of the farm. When read across, it appears that incomes increase with size of farm, regardless of yields. If the table is read diagonally across from the upper left to the lower right, it appears that large farms with good crop yields made $1,261, compared with a loss of $110 for small farms with poor yields.

---

[4] The numbers of farms in groups were 352, 284, and 271.

In the first column of table 3, only farms of less than 215 units are considered; in the second column, only farms of 215 to 344 units; and in the third column, only farms of 345 units or more. Thus, in each column, the size of farms is *held constant* in order to study the relation of yield to income when size remains the same. The first column of table 3 shows that *for small farms* incomes rose from −$110 to +$202 as crop yields improved.

When the table is read across, size of farm fluctuated, while yields were held constant at three different levels. When crop yields were constant at 110 or more, incomes increased from $202 to $1,261.

A two-way table is the simplest and frequently the most effective way of studying the "multiple" relationship of two independent variables to one dependent variable.[5]

THREE-WAY TABULAR ANALYSIS

A three-way table shows the relationship of three independent variables to the dependent variable. The farms were first divided into two approximately equal groups on the basis of "poor" and "good" yields, the first independent variable. These groups were then subdivided on the basis of size of farm and again subdivided on the basis of labor efficiency.[6] There were eight groups containing from 45 to 181 farms each,[7] for which the totals and averages for income were obtained.

The eight averages were arranged in an orderly manner in table 4. A three-way table is much more difficult to interpret than a two-way table because there are more possible comparisons. For instance, one might be interested in studying the effect: (1) of size on income, with yields and efficiency held constant; (2) of efficiency on income, with yields and size held constant; and (3) of yields on income, with size and efficiency held constant.

---

[5] A multiple relationship is one involving two or more independent variables.

[6] The order of grouping and subgrouping does not affect the numbers in the subgroups or the averages.

[7] Although each independent variable was divided into two approximately equal groups, there was considerable variation in the number of farms within subgroups. This was due to the presence of an interrelationship. There were 45 farms with "low" efficiency, "large" size (and "poor" yields). "Low" efficiency on "large" farms was not common.

There were 181 farms with "low" efficiency, "small" size (and "poor" yields). Low efficiency on small farms seemed to be the rule. In other words, there was an interrelationship between two independent variables, size and efficiency. When such interrelationships exist, the size of subgroups will not be uniform.

If a student were interested in the first relationship, size and income, he would compare the average incomes in the third column with the corresponding averages in the fourth. Except when crops were poor and efficiency was low, incomes increased with size of farm. Size of farm is an important factor affecting income, but the nature of the effect of size depends on whether other factors are favorable. Since poor yields and low efficiency make for losses, large farms with such conditions would lose more than small farms.

If the student were interested in the second relationship, the effect of efficiency on income, he would compare the average incomes in the first row with the corresponding averages in the second; and those in the third, with those in the fourth. With yield and size held constant at poor and small, respectively, income rose from −$119 to $384 as efficiency increased from low to high. A similar comparison for the other three pairs of incomes shows the same general relationship.

If the student were interested in the third relationship, the effect of yields on income, he would compare the average incomes in the first row with the corresponding averages in the third; and those in the second, with those in the fourth. With efficiency and size held constant at high and large, income rose from $592 to $1,139 with better crop yields.

TABLE 4.—THREE-WAY TABLE

RELATION OF CROP YIELDS, SIZE OF FARM, AND LABOR EFFICIENCY TO INCOME, 907 NEW YORK FARMS, 1927

Crop yields	Efficiency	Size of farm	
		Small	Large
		Income	*Income*
Poor	low	$ −119	$ −271
Poor	high	+384	+592
Good	low	+101	+232
Good	high	+361	+1,139

FOUR-WAY AND HIGHER-ORDER TABULAR ANALYSIS

There is no limit to the number of variables in higher-order tabular analysis except the number of observations and the number of variables which are related. With each additional subclassification, the number of groups is increased, and the number of observations in each group is thereby decreased. Unless the total number of observations is very large, the numbers in each group are likely to become too few for reliable averages. In the three-way table, the smallest subgroup contained 45 farms. In a four-way table, which would involve further subgrouping, the smallest number would probably be 15 to 20 farms. In a five-way table, the smallest number might be reduced to 5 or 6.

The difficulty of interpreting tables increases "geometrically" with the number of variables involved. Fourth- and higher-order tables become so complicated that it is almost impossible for the human mind to grasp the significance of all the many relationships shown. Furthermore, four-way and higher-order analyses are not so useful as one-, two-, and three-way analyses, because in most problems the greater part of the variability is due to the effects of only one, two, or three independent variables.

LINEAR RELATIONSHIPS

In tabular analysis, one can quickly observe by inspection whether any relationship exists and, if so, the direction of the relationship. With a few simple calculations or with a simple graph, it is possible to determine whether the relationship is linear or curvilinear. A relationship is linear when each unit increase in the independent variable is accompanied by a constant increase in the dependent variable throughout the range of the data. For example, an increase of 10 in X_2 might always be accompanied by an increase of 6 in X_1, regardless of how large or how small X_2 became. This would be called a linear relationship.

TABLE 5.—TABULAR ANALYSIS OF A SIMPLE LINEAR RELATIONSHIP

Relation of Size of Business to Operating Costs of 173 Cooperative Creameries,* Minnesota, 1934

Butter made, pounds	Cost per pound, cents
Less than 125,000	3.55
125,000 to 249,999	3.21
250,000 to 374,999	2.91
375,000 to 499,999	2.62
500,000 to 624,999	2.34
625,000 and over	2.13
All groups	2.65

* Koller, E. F., and Jesness, O. B., Minnesota Cooperative Creameries, University of Minnesota, Agricultural Experiment Station, Bulletin 333, p. 61, September 1937. Ordinarily, tables give the number of items in each group. The numbers of creameries from small to large groups were 9, 53, 50, 32, 15, and 14.

The effect of size of business on the cost of making a pound of butter illustrates a simple linear relationship (table 5). The cost per pound of butter, the dependent variable, is related to the size of business, the independent variable. Since the cost declines with increasing volume, the relationship is a negative one.

The relationship is also linear. Each increase of 125,000 pounds in size of business decreased the cost about 0.30 cent per pound. The difference between the first two groups was 0.34 cent; the second and third, 0.30; third and fourth, 0.29; and fourth and fifth, 0.28.

The linearity of the relationship becomes more apparent in graphic form (figure 1). The approximately constant rate of decrease in cost is shown by an approximately straight line. Since the rate of decrease in cost was practically constant for the whole relationship, it may be approximated by a single value. The difference between the first and fifth groups was 1.21 cents. Since the difference in size of business was 500,000 pounds of butter, the rate of decrease in cost was 0.30 cent per pound per each increase of 125,000 pounds in production [1.21 ÷ (500,000 ÷ 125,000) = 0.3025], or a decrease of 0.0000024 cent per pound with every increase of 1 pound in production (1.21 ÷ 500,000 = 0.00000242).

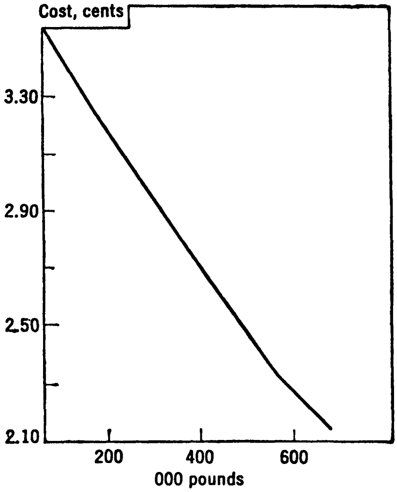


FIGURE 1.—LINEAR RELATIONSHIP

Relation of Size of Business to the Cost of Making a Pound of Butter

As size of business increased, costs declined at a fairly constant rate. Such a relationship is termed *linear* because it follows the pattern of a straight line.

CURVILINEAR RELATIONSHIPS

The relation between the age and yields of peach trees illustrates the tabular analysis of a simple curvilinear relationship (table 6). The technique is no different from that for the linear relationship in table 5. The difference is in the relationship itself.

As peach trees grew from 3 to 13–14 years of age, yields *increased* from 0.64 to 1.76 bushels per tree. From 13–14 to 23–24 years, yields *decreased* from 1.76 to 0.86 bushel per tree (table 6). In other words, as trees grew older, yields increased until the trees reached the age of about 13–14 years and decreased thereafter. Because the change in yields with each increase in age was not constant or even in the same direction, the relationship was curvilinear rather than linear.

A curvilinear relationship may be most vividly shown in graphic form. The average yields for different ages of trees were plotted and connected with a solid line (figure 2). In general, as age increased, yield increased at the younger ages and decreased at the older ones. The relationship may be generalized by the smooth, broken curve which first increases at a decreasing rate, then reaches a maximum, and finally decreases at an increasing rate.

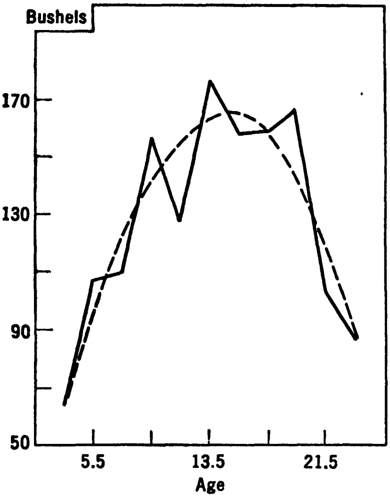


FIGURE 2.—CURVILINEAR RELATIONSHIP

AGE AND YIELD OF PEACH TREE

As age increased, yields increased until trees were about 13 to 14 years of age, and then declined. Such a relationship is termed *curvilinear* because it follows the pattern of a curve.

TABLE 6.—TABULAR ANALYSIS OF A SIMPLE CURVILINEAR RELATIONSHIP

RELATION OF AGE OF PEACH TREES TO YIELD PER TREE*

Age, years	Yield, bushels
3- 4	0.64
5- 6	1.07
7- 8	1.10
9-10	1.56
11-12	1.27
13-14	1.76
15-16	1.58
17-18	1.59
19-20	1.67
21-22	1.03
23-24	0.86

*DeGraff, H. F., The Influence of Soil on Peach Yields and Peach Tree Mortality, Farm Economics, No. 104, p. 2535, December, 1937. Based on records for the 11-year period 1926–1936, Newfane-Olcott Area, Niagara County, New York.

ABSENCE OF RELATIONSHIPS

A common tendency is to ignore those analyses which show no relationship. Frequently, it is just as important, however, to know that no relationship exists among certain variables as it is to know that it exists among others.

ADDITIVE RELATIONSHIPS

The effects of two independent variables on a dependent variable have already been studied by two-way tabular analysis (table 3). For the purpose of studying factors affecting labor income, 520 farms were divided into three groups on the basis of an independent variable, crop yields (table 7). Each of these three groups was subdivided into three subgroups on the basis of the second independent variable, labor efficiency. The average of the dependent variable, income, was calculated for each of the nine combinations. The lowest income, –$87, was obtained for farms on which both crop yields and labor efficiency were low. Conversely, the highest income, $1,375, was made when both yields and efficiency were high.

TABLE 7.—TWO-WAY TABULAR ANALYSIS OF AN ADDITIVE RELATIONSHIP, WITH TWO INDEPENDENT VARIABLES

RELATION OF LABOR EFFICIENCY PER MAN AND CROP YIELDS TO INCOME, 520 FARMS, NEW YORK, 1908

Yields,* index	Labor efficiency,* units		
	Less than 200	200–249	250 or more
	Income	*Income*	*Income*
Less than 85	$–87	$ 210	$ 565
85 to 114	283	621	946
115 or more	675	1,038	1,375

* The average indexes of crop yields for the three groups were 73, 100, and 129. The average indexes of labor efficiency for the three groups were 146, 218, and 309.

The relationship of yields to income was "additive." A relationship between one independent and the dependent variable is called additive when that relationship is the same regardless of the sizes of the other independent variables. The effect of yield on income was the same regardless of whether efficiency was high, medium, or low.

When efficiency was as low as 200 work units per man, increased yields resulted in higher incomes (table 7). The difference in incomes for farms with high and low yields was $762 [675 – (–87) = 762]. When efficiency was average, the difference in incomes with high and low yields was $828 (1,038 – 210 = 828). When efficiency was high, this difference was $810. Regardless of labor efficiency, an increase in crop yields resulted in about the same number of dollars increase in incomes.

These increases, which averaged $800, represented the *net* increase in income with a change in yields from poor to good, with the effects of low and high efficiency held constant. This is not the same as the total effect of yields which one would obtain by grouping by yields alone and not considering efficiency.

The average increase[8] due to yields was \$800 [(762 + 828 + 810) ÷ 3 = 800]. This increase accompanied a change of 56 points in the index of crop yields (129 − 73 = 56). The *average effect* of each unit change in yields was \$14.29 (800 ÷ 56 = 14.29).

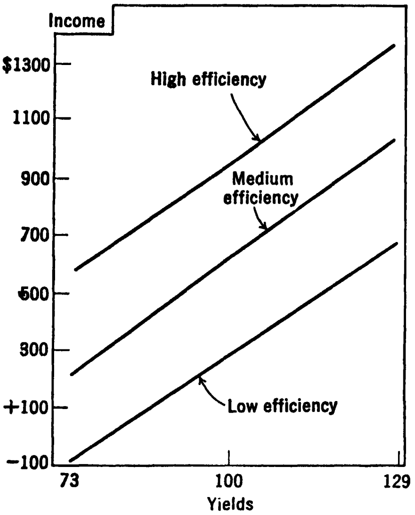


FIGURE 3.—AN ADDITIVE RELATIONSHIP

RELATION OF YIELDS TO INCOME FOR FARMS WITH "HIGH," "MEDIUM," AND "LOW" LABOR EFFICIENCY

The effect of yields on incomes was about the same regardless of efficiency. All three lines sloped upward, and the rates of change were about the same. Such a relationship is termed *additive*.

The nature of an additive relationship may be more clearly observed from its graphic form. The relationships between yields and income for high, medium, and low efficiency were plotted as three separate lines (figure 3). The three lines all increased at about the same rate; that is, the lines were approximately parallel. This indicated that, as yields became better, incomes rose by about the same amount regardless of whether efficiency was high, medium, or low.

When crop yields were low, less than 85, the difference between incomes on farms with low and high efficiency was \$652 [565 − (−87) = 652]. For moderate and high yields, these differences were \$663 and \$700 (table 7). Regardless of crop yields, an increase in labor efficiency resulted in about the same increases in incomes. The average increase due to efficiency was \$672 [(652 + 663 + 700) ÷ 3 = 672]. This increase accompanied a change of 163 work units per man (309 − 146 = 163). The *average effect* of each unit increase in efficiency was \$4.12 (672 ÷ 163 = 4.12). The effect of efficiency on income for three different yields could be shown in a chart similar to figure 3. The three lines would be approximately parallel.

---

[8] This average increase is an unweighted difference. Methods of weighting differences by the number of observations are given by Harper, F. A., Analyzing Data for Relationships, Cornell University Agricultural Experiment Station, Memoir 231, p. 10, June 1940.

In figure 3, the effect of efficiency is shown by the differences between the three lines. The middle line is above the lowest line because of a difference in efficiency. The fact that the middle line is about the same distance above the lowest line throughout their lengths indicates that the effect of efficiency was always the same regardless of yields. Thus, in figure 3, the effect of yield is shown by the *slopes* of the lines; and the effect of efficiency, by the *levels* of the lines.

The fact that the lines all have the same slope indicates that the effect of yields was constant for all levels of efficiency. Also, the fact that all the lines were parallel indicates that the effect of efficiency was constant for all levels of yield. The mathematically adept will note that whenever all three lines have the same slope they *must* be parallel, and vice versa. This means that, when the effects of yields are constant for all levels of efficiency, the effects of efficiency must be constant for all levels of yields. In other words, if the effect of yields is independent of efficiency, so also is the effect of efficiency independent of yields.

Figure 3 gives a clue as to why such relationships are called additive. The average income for poor yields and low efficiency was −$87, the lowest point on the lowest line of figure 3. Since the effects of both yields and efficiency are constant for all values of each other, the income for any combination of yields and efficiency may be obtained by adding to −$87 the average effect of yield and the average effect of efficiency.[9]

All the points to the right on the lowest line of figure 3 measure the additive effects of yields. These effects are termed additive because they are the same as the *average* effects of yields measured on all three lines. In fact, they should have been termed average effects instead of additive effects.

As one moves from the lowest point of the lowest line, −$87, to the lowest point of the middle line, $210, and of the upper line, $565, the differences between these low points measure the additive effects of efficiency (figure 3 and table 7). These effects are termed additive because they are the same as the *average* effects of efficiency measured by the differences between lines at three different points.

If one moves from the lowest point of the lowest line, −$87, to the center of the middle line, $621, and to the highest point of the upper line, $1,375, the differences between these points measure the additive effects of both yields and efficiency. Again, these effects are called additive because the effects of yields and of efficiency between any two points are always about the same as the average effects of yields and efficiency for the whole problem, $14.29 and $4.12 per unit, respectively.

---

[9] The average increase due to high over low efficiency was $672; and that due to good over poor yields, $800.

Relationships are said to be additive when the dependent variable for any combination of two independent variables can be accurately determined by adding their average effects.

From the two-way tabular method, the following facts were learned: The relationships of both independent variables to the dependent variable were positive, linear, and additive; the net rate of increase in income with each unit increase in yields was $14.29, and in labor efficiency, $4.12. The presence of a multiple relationship was unquestionable.

JOINT RELATIONSHIPS

When the effect of an independent variable is not constant for all values of another independent variable, the relationships are not additive. They are joint.

TABLE 8.—TWO-WAY TABULAR ANALYSIS OF A JOINT RELATIONSHIP, WITH TWO INDEPENDENT VARIABLES

Relation of Yields and Size of Business to Income, 620 Tobacco Farms, Virginia,* 1933

Yields, index	Size of business, units		
	Less than 350	350–599	600 or more
	Income	*Income*	*Income*
Less than 85	$−254	$−329	$−564
85 to 109	−126	−135	−118
110 or more	− 83	114	474

* Underwood, F. L., Flue-cured Tobacco Farm Management, Virginia Agricultural Experiment Station, Technical Bulletin 64, p. 222, January 1939. Total productive man-work units were used as a measure of size.

The two-way tabular method may be applied to joint as well as to additive relationships. Both independent variables, crop yields and size of business for Virginia tobacco farms, were related to the dependent variable, income (table 8). The effect of size of business on income, with yields held constant, can be observed by reading the rows of table 8 from left to right. This relationship was sometimes positive and sometimes negative. When yields were low, the relationship was negative;

and large farms received $310 *less* than small farms [−564 − (−254) = −310]. When yields were average, large and small farms lost about the same amounts. When yields were high, large farms returned $557 *more* income than small farms (table 8). Apparently, the effect of size on income depended on yields. Size and yields thus were jointly related to income.

When the effect of one independent on the dependent variable changes with different values of the other independent variable, the relationships are said to be joint.

The nature of joint relationships can be shown graphically (figure 4). The three lines represent the relationships between size and income for three different yields. The three lines sloped in three different directions. This indicated that, as size increased, income might decline rapidly, stay the same, or rise rapidly, depending on whether yields were poor, medium, or good.

The *average* effect of large over small size was $85 [(−310 + 8 + 557) ÷ 3 = 85]. However, the average effect does not tell the whole story here because the actual effect of size for any given yield was never $85.

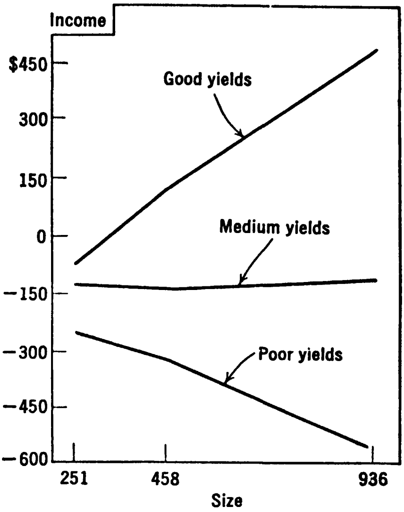


FIGURE 4.—A JOINT RELATIONSHIP

Relation of Yields and Size of Business to Incomes with "Good," "Medium," and "Poor" Yields

The effect of size on income is different with varying crop yields. The slopes of the lines and the rates of change were all different. Such a relationship is termed *joint*.

The relation of yield to income, with size of business held constant, was always positive, regardless of whether farms were large or small (table 8). However, increases in crop yields raised incomes more on large farms than on small ones. On small farms, the increase in incomes was $171 [−83 − (−254) = 171]. For average-sized and large farms, the corresponding increases were $443 and $1,038. Because these increases were all greatly different and therefore not the same as their average, the relationship of yields to income was also said to be joint.

When one independent variable is jointly related with a second independent variable in its effect on the dependent, so is the second independent jointly related with the first independent variable. Since size was jointly related with yields in its effect on income, so yield was jointly related with size in its effect on income.

The joint relation changed the effects of size and of yields somewhat differently. As yields became larger, the effect of size changed from a large decrease to a large increase. On the other hand, the effect of yields changed from a small increase to a very large increase.

The joint effects of size and yields as observed in table 8 are not unusual. Success or failure in farming depends partly on yields. Costs are about the same regardless of yields, and, consequently, high-yielding farms usually return a surplus; and low-yielding farms, often a deficit.

The amount of surplus or deficit, of course, depends partly on how good or how poor the yields are, but also on the size of the farm. If a farmer loses money because of poor yields, he can lose much more on a large farm than on a small one. Conversely, if he profits from good yields, he can make more on a large farm.

With the two-way tabular method, the following facts were learned concerning the effects of size of business and yields on income from tobacco farms: The relationships of these independent variables to income were both positive and negative, approximately linear, and joint. The increases in income with size were −$310, +$8, and $557; and with crop yields, +$171, +$443, and $1,038. The presence of a joint relationship was unquestionable.

If a relationship is additive, this fact is apparent in the final averages. Likewise, if the relationship is joint, this fact will appear. In either event, the method of grouping the observations and obtaining totals and averages is exactly the same. A knowledge of the nature of relationships comes after the work has been completed.

NON-NUMERICAL INDEPENDENT VARIABLES

Thus far, all variables have been measured in numerical terms. It is often desirable to analyze the relation between a dependent variable numerically described, such as price or income, and one or more independent variables qualitatively described, such as type of soil, sex, race, color, breed, or grade. Tabular analysis is as adaptable to non-numerical as to numerical independent variables.

The effect of grades on the price of steers illustrates the tabular analysis where the independent variable is non-numerical (table 9). The independent variable, quality, is described not by numbers, but by adjectives. The relation between the quality and price of cattle was positive. The differences in price between successive grades from common to choice became less and less with each increase in grade. This might lead one to the conclusion that the relationship was curvilinear. In reality, there is nothing in table 9 to indicate the pattern of the relationship. There is one grade difference between common and medium and also between medium and good. This does not mean that the two differences are the same in terms of some numerical measure of quality. For this reason, a unit rate of change cannot be calculated for a relationship in which the independent variable is non-numerical, and whether the relationship is linear or curvilinear cannot be determined.

TABLE 9.—ONE-WAY TABULAR ANALYSIS OF A RELATIONSHIP WITH A NON-NUMERICAL INDEPENDENT VARIABLE

RELATION OF GRADE TO THE PRICE OF CATTLE* PER 100 POUNDS, CHICAGO, 1939

Grade	Price
Common	$ 7.51
Medium	8.77
Good	9.81
Choice	10.48

* Jordan, E. M., Livestock, Meats, and Wool Market Statistics, 1939, mimeographed report of the United States Department of Agriculture, p. 54, May 1940.

Multiple relationships with two non-numerical independent variables can also be analyzed by tabulation (table 10). The value of farm land varied with the type of road and class of land. The highest valuations were found for good land on concrete roads. When land was poor, value was not associated with the type of road, whereas, for good land, farms on concrete roads were worth about one-third more than those on dirt roads.

Tabular analysis with three, four, or more non-numerical independent variables is also possible. The relation of season, type of store, and size

TABLE 10.—TWO-WAY TABULAR ANALYSIS WITH TWO NON-NUMERICAL INDEPENDENT VARIABLES

RELATION OF ROADS AND LAND CLASS TO THE VALUE OF FARM LAND, CLINTON COUNTY,* NEW YORK, 1934

Type of road	Land class		
	Poor	Fair	Good
	Value land	*Value land*	*Value land*
Dirt or gravel	$15	$18	$41
Macadam	11	25	46
Concrete	17	39	56

* White, O. H., An Economic Study of Land Utilization in Clinton County, New York, Cornell University Agricultural Experiment Station, Bulletin 689, p. 45, April 1938.

TABLE 11.—THREE-WAY TABULAR ANALYSIS WITH THREE NON-NUMERICAL INDEPENDENT VARIABLES

RELATION OF SEASON, TYPE OF STORE, AND SIZE OF CONTAINER TO WEEKLY SALES OF EVAPORATED MILK,* ROCHESTER, NEW YORK, 1931

Size of can	Winter		Summer	
	Independent	Chain	Independent	Chain
	Cans milk sold	*Cans milk sold*	*Cans milk sold*	*Cans milk sold*
Small.......	60	162	62	208
Large.......	75	243	65	266

* Mumford, H. W., The Sale of Milk and Cream through Retail Grocery Stores in Rochester, New York, Farm Economics, No. 80, p. 1911, May 1933.

of container to sales of evaporated milk involves three non-numerical independent variables (table 11). Sales of evaporated milk were greater in chain than in independent stores, greater in large cans than in small ones, and usually greater in summer than in winter.

Of the three factors related to the retail price of potatoes, type of store and income area were non-numerical, but grade was numerical (table 12). Type of store and income were not related to price. Grade may have affected price slightly, but the relationship was doubtful.

TABLE 12.—THREE-WAY TABULAR ANALYSIS WITH TWO NON-NUMERICAL AND ONE NUMERICAL INDEPENDENT VARIABLES

RELATION OF GRADE, TYPE OF STORE, AND INCOME AREA TO THE RETAIL PRICE OF POTATOES PER PECK,* ROCHESTER, NEW YORK, 1936–37

Grade, per cent U. S. No. 1	Independent stores		Chain stores	
	High-income area	Low-income area	High-income area	Low-income area
	Price potatoes	*Price potatoes*	*Price potatoes*	*Price potatoes*
75–89...........	35.7¢	33.4¢	35.0¢	34.3¢
90 or more......	35.7	35.5	36.4	36.6

* Findlen, P. J., Relation of Income to Grade of Potatoes Sold in Rochester and Buffalo, New York, Farm Economics, No. 110, p. 2684, November-December 1938.

NON-NUMERICAL DEPENDENT VARIABLES

When the dependent variable is numerical, relationships are usually studied by averaging it for different classifications of the independent variables (tables 1 to 12). When the dependent variable is non-numerical, its average cannot be obtained, and some other method of analyzing the relationship must be used. A common method is to classify the data first according to the independent and then subclassify according to the dependent variable, or vice versa. Then, the numbers of observations falling in the different combinations are counted.

TABLE 13.—TWO-WAY TABULAR ANALYSIS WITH NON-NUMERICAL INDEPENDENT AND DEPENDENT VARIABLES

RELATION OF CONDITION OF THE BARN TO THE CONDITION OF THE HOUSE, REFORESTATION AREAS,* NEW YORK STATE, 1935

Condition of house	Condition of barn				
	Gone	Falling	Poor	Fair	Good
	Number farms	*Number farms*	*Number farms*	*Number farms*	*Number farms*
Gone	51	5	10	4	2
Falling	8	20	6	1	0
Poor	8	8	29	6	1
Fair	6	2	4	23	11
Good	0	2	2	6	6
Total	73	37	51	40	20

* La Mont, T. E., State Reforestation in Two New York Counties, Cornell University Agricultural Experiment Station, Bulletin 712, p. 13, February 1939.

On New York farms purchased by the state for reforestation, the condition of the house was related to the condition of the barn (table 13). On 73 farms where the barn was gone, 51 of the houses were also gone; 8, falling; 8, poor; 6, fair; and none were good (table 13). On the 40 farms where the barn was fair, only 4 houses were gone, and 29 were fair or good. Apparently, the condition of the house was related to the condition of the barn. There might be some difference of opinion as to which should be considered the dependent variable. If the condition of the barn were dependent on the house, table 13 might be analyzed row by row instead of column by column.

Table 13 is really a two-way frequency distribution in which the variables are non-numerical.

Frequently, the number of farms would be changed to percentages of totals. For example, when the barn was gone, 70 per cent of the houses were also gone (51 ÷ 73 = 0.70). When the barn was fair, the house was gone for 10 per cent of the farms (4 ÷ 40 = 0.10).

TABLE 14.—TWO-WAY TABULAR ANALYSIS WITH A NUMERICAL INDEPENDENT AND NON-NUMERICAL DEPENDENT VARIABLE

RELATION OF SIZE OF FARM TO THE CROPPING SYSTEM,* IRRIGATED FARMS, WYOMING

Size of farm, acres	Kind of crop				
	Sugar beets	Beans	Alfalfa	Small grain	Other
	Per cent of crop acres	*Per cent of crop acres*	*Per cent of crop acres*	*Per cent of crop acres*	*Per cent of crop acres*
20 or less	15	21	27	25	12
21- 40	10	23	39	25	3
41- 60	16	27	31	24	2
61- 80	14	38	27	19	2
81-100	14	34	36	15	1
101 and over	11	27	31	29	2

* Vass, A. F., and Pearson, H., Economic Studies of Irrigated Farms in Big Horn County, Wyoming Agricultural Experiment Station, Bulletin No. 205, p. 72, May 1935.

When the independent variable is numerical and the dependent variable is non-numerical, the same method of analysis is usually employed. A study of irrigated farms included a numerical independent variable, size of farm, and a non-numerical dependent variable, kind of crops. The object of the tabulation was to determine whether the size of farm was related to the kind of crops grown. Apparently, as farms became large, the importance of beans increased and grain decreased until the farms were over 100 acres in size (table 14). On most farms, only 1 to 3 per cent of the land was not in either sugar beets, beans, alfalfa, or small grain. However, on very small farms, 12 per cent of the land was in other crops.

Two- and three-way frequency distributions are relatively inefficient methods of analyzing and showing relationships.[10] When the dependent variable is non-numerical, however, other methods are not adaptable. When the independent variable is numerical, some students attempt to simplify the problem by grouping data according to the dependent variable and tabulating the independent.

---

[10] A relationship shown by a two-way frequency distribution can be shown by a one-way tabulation when one of the variables is numerical. If each variable had five classifications, there would be 25 numbers in the frequency distribution and only 5 averages in the tabulation. A relation shown by 5 numbers is usually more effective than one shown by 25 numbers.

HOLDING INTERRELATED VARIABLES CONSTANT

With two-way and higher-order tabular analysis, it is possible to study the effect of one independent variable "holding other variables constant." The effect of the independent variable, size of herd, on the dependent variable, income, with labor efficiency held constant, can be studied by comparing average incomes from different-sized herds on farms with the same labor efficiency (table 15). When herds were small and labor efficiency low, income was $190, compared with $406 when herds were small and efficiency was high. The difference, $216, might be ascribed to the difference of 11 tons of milk per man. However, labor efficiency and size of herd are interrelated. In a strict sense, labor efficiency is not an independent variable, but itself depends on the size of herd. Of the herds classified as small, the less efficient farms averaged 9 and the more efficient farms 13 cows. The classification really does not hold the number of cows completely constant. The difference of $216 is only partly due to the difference in efficiency. Some of the difference in income is probably due to the difference of 4 cows in size of herds.

TABLE 15.—TWO-WAY TABULAR ANALYSIS WITH INTERRELATED INDEPENDENT VARIABLES

RELATION OF SIZE OF HERD AND LABOR EFFICIENCY TO INCOME

Size of herd, number of cows		Labor efficiency, tons of milk per man		Income, average
Group	Average	Group	Average	
Small	9	low	20	$190
Small	13	high	31	406
Large	19	low	27	384
Large	27	high	36	692

Likewise, when the number of cows was large, the difference of $308 in income between farms high or low in efficiency was partly due to the difference, 8, in the number of cows.

The difference in incomes on large and small farms with low efficiency appeared to be $194 (384 − 190 = 194). However, part of this difference was probably due to the difference between 20 and 27 tons of milk per man.

The above example is an illustration of an extremely high degree of interrelationship. Most interrelationships are much less marked. Nevertheless, the failure of the tabular method to hold completely constant the effect of interrelated variables is a shortcoming. This failure can be minimized by increasing the number of classifications for both independent variables. Many students who recognize this shortcoming take pains to obtain group averages for not only the dependent but also the independent variables (as in table 15). From the averages of the independent variables, disturbing interrelationships can be detected.

CHARACTERISTICS

Tabular analysis is a method of analyzing relationships. Relationships are the most important problems of life and, consequently, the statistician's most important problem. The great mass of scientific workers in all fields use tabular analysis almost to the exclusion of all other methods. However, most textbooks ignore the tabular method of analyzing relationships or merely give it brief comment.[11]

The tabular method shows whether any relationship exists. This is indicated by the consistency with which the dependent variable fluctuates with changes in the independent variable. It shows whether the relationships are positive or negative. It indicates whether the relationships are linear or curvilinear. It shows the rates of change for linear relationships and the nature of the curves for curvilinear relationships. The method indicates whether the relationships are joint or additive. If they are joint, the nature of the joint relationships is revealed.

When the number of observations is small, tabular analysis has an important defect. The number of data represented by group averages is too small to give reliable results. The number of observations in any one group depends not only on the total number but also on the number of groups and the distribution within groups. The reliability of a group average depends not only on the number of items but also on the variability in the data and the degree of relationship. Nevertheless, the possibilities of tabular analysis in showing relationships accurately, completely, and in adequate detail are directly limited by the total number of observations. For scanty data, this fault outweighs all the important advantages of tabular analysis.

---

[11] Most textbooks discuss in considerable detail methods of constructing tables and ignore tabulation as a method of analyzing relationships.

Bowley, A. L., Elements of Statistics, Fourth Edition, p. 62, 1920, points out that tabulation is a method of showing correlation, the correspondence in the occurrence of two sets of phenomena.

Jones, D. C., A First Course in Statistics, p. 18, 1921, points out that tabulation is a very useful method of studying correlation. He proceeds no farther with tabulation, but devotes several chapters to correlation.

Ezekiel, M., Methods of Correlation Analysis, 1941, recognizes the problem and has two short chapters on tabulation. Its use is discussed mostly on the basis of the large number of observations needed, and the fact that it gives no exact index of the closeness of association.

The subgrouping device of the tabular method does not always hold the effect of independent variables constant. However, this limitation is serious only when there are marked interrelationships between independent variables and when the number of classes is small.

Tabular analysis is somewhat "wasteful" of data. Because of interrelationships, the number of observations in the different subgroups of a table are not the same. In comparing an average of 20 items with an average of 5 items, the reliability of the comparison is limited by the smaller group.

The usefulness of higher-order tables is somewhat limited by their complexity. One-way tables are easy to analyze. Two-way tables are somewhat more difficult. Three- and four-way tables require much more time than most persons are willing to give. In such tables, the number of figures to be compared, the number of relationships, and their joint effects multiply. A three-way table with 3 different values for each independent variable contains 27 different values of the dependent variable and 3 general relationships, as follows:

X_2 and X_1 (27 comparisons)
X_3 and X_1 (27 comparisons)
X_4 and X_1 (27 comparisons)

It also contains 4 joint relationships each of which can be examined in many different comparisons. The 7 different relationships could be examined in 100 to 200 individual comparisons. This multiplicity of possible comparisons makes three- and four-way tables difficult to interpret.

A large amount of judgment and common sense is required for using and interpreting the results of any statistical method, including tabular analysis.

The tabular method has several important advantages over other methods of showing relationships.

From the standpoint of simplicity of the method of analysis, tabulation has overwhelming advantages. Simplicity of method is extremely important for at least two reasons:

1. It saves time.
2. It makes possible the factual study of science's most important

problem—the analysis of relationships—by the great mass of research workers in many fields. The great majority of scientific workers have neither the time nor the inclination to familiarize themselves with more complicated statistical methods. In fact, if they had, they probably would not have accomplished much in their own fields. Scientists can visualize a simple process, and, because they can, they have confidence in it. They use it because they understand it. Most research workers accept the simpler techniques, provided that they are satisfactory.

From the standpoint of simplicity of presentation, tabulation has an overwhelming advantage. Tabulation shows relationships in terms commonly understood by all classes of readers. The rank and file of laymen understand a table but are confused by more complicated statistical results.

Tabulation is a flexible type of analysis. The nature of the relationship is not assumed at the outset. The technique is the same whether the relationships are linear or curvilinear, additive or joint. The discovery of these characteristics comes in the interpretation after the tabulations have been made.

When one or more variables are non-numerical, tabulation methods are just as simple and effective as when the variables are numerical.

CHAPTER 9

CORRELATION

Farmers generally recognize that more labor is required to harvest large crops than small ones. On 16 farms in 1938, it was found that only 7.2 hours were required to harvest small crops; 10.6 hours, to harvest average crops; and 16.0 hours, to harvest large crops (table 1). This simple statistical analysis verifies with a considerable degree of accuracy the observations of farmers.

TABLE 1.—RELATION OF YIELD OF ALFALFA TO HOURS OF MAN LABOR REQUIRED TO HARVEST AN ACRE

16 New York Farms, 1938

Yield, tons	Hours
1.4 to 2.2	7.2
2.3 to 3.0	10.6
3.1 to 4.1	16.0

This problem of association is one of the most important problems in life. It confronts all people in all walks of life from birth until death, in their individual and collective activities. Some relationships are very simple; others involve varying degrees of complexity. Most of our knowledge concerning these relationships is based on experience, because it is impossible for the layman to make statistical analyses and determine definite relationships. For those relationships which appear to be worthy of analysis, the majority of scientists follow tabular analysis,[1] as illustrated in table 1. The statistician, however, has developed methods of expressing such relationships in one number instead of several. This is the problem of correlation.

The correlation coefficient is an attempt to summarize in one number the amount of relationship existing between two things. Many students have difficulty in understanding correlation because of the complexity of the methods used and because of the difficulty in interpreting the result. The first problem of the student is to understand the principles involved in correlation; and the second, its calculation.

Association and the principle of correlation may be studied graphically. Observations may be arranged on a graph according to the way they vary in respect to two characteristics. In a study of the relation of yield to labor required in harvesting an acre of alfalfa, one observation is one farm in one year. The data for one farm, of course, consist of the average yield and the hours per acre required to harvest the crop. The problem is to measure the manner and extent that yield varies with labor from farm to farm. In figure 1, the vertical scale represents labor; and the horizontal scale, the yield. Each dot is an observation whose location is determined by the two characteristics yield and labor per acre.[2] Even the most casual observer will note that, as the dots are placed farther to the right, that is, in the direction of the higher yields, they tend also to fall higher on the graph, in the direction of more labor per acre. Hence, a relationship between yield and labor is indicated; as yield increases, more labor is required to harvest the crop.

---

[1] Tabulation analysis is discussed in detail in chapters 8 and 15, pages 120 and 264.

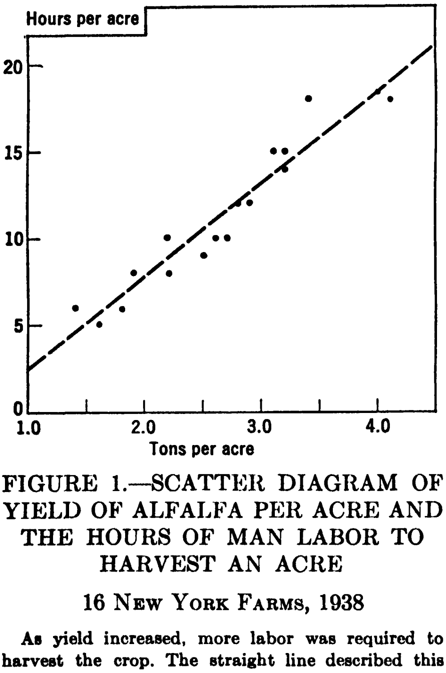


FIGURE 1.—SCATTER DIAGRAM OF YIELD OF ALFALFA PER ACRE AND THE HOURS OF MAN LABOR TO HARVEST AN ACRE

16 New York Farms, 1938

As yield increased, more labor was required to harvest the crop. The straight line described this tendency.

This relationship may be generalized and described by a straight line drawn through the dots in such a way that it may be said to fit better than any other straight line that could be drawn.[3] Just as any single variable may be described by its arithmetic mean, so may the relationship between two variables be described by the "line of relationship."

Of course, not all the individual items equal the arithmetic mean. Neither do they all fall on the line of relationship. The amount that the items deviate from the arithmetic mean may be shown by short solid lines perpendicular to the horizontal line representing the arithmetic mean (figure 2). The squares formed by the short solid and dashed lines represent the deviations squared. The sum of these squared deviations, which may be written $\Sigma\boxed{1}$, is a measure of the degree to which the arithmetic mean *does not* accurately describe the data.

---

[2] Such a chart is commonly called a scatter diagram.

[3] Several bases for judging goodness of fit could be used. In practice, the line which applies here is that determined by the method of least squares, $Y = -3.0203 + 5.3924X$.

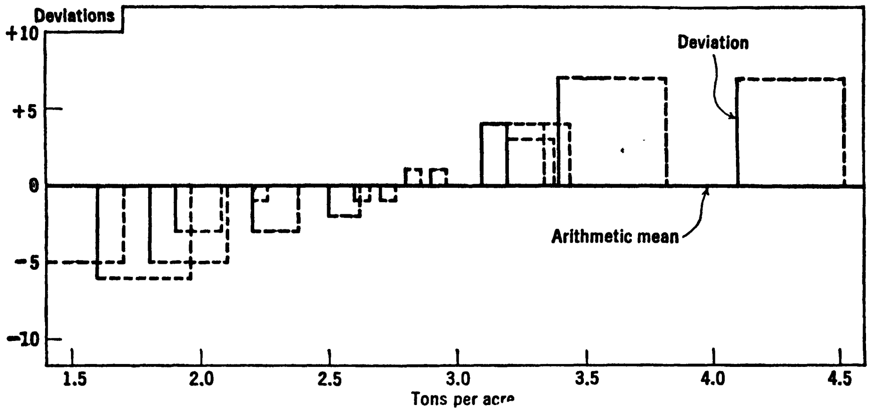


FIGURE 2.—DEVIATIONS AND SQUARES OF DEVIATIONS FROM ARITHMETIC MEAN

YIELD OF ALFALFA AND HOURS TO HARVEST AN ACRE ON 16 NEW YORK FARMS, 1938

The amounts that hours of labor deviate from the arithmetic mean are shown by the short, solid, vertical lines. These solid lines are perpendicular to a horizontal line representing the mean. The squares of these deviations are represented by the areas enclosed by the solid and broken lines.

In the above figure the single variable, labor, was described by a horizontal line. Labor may be also described by the line of relationship between labor and yield. This line of relationship seems to describe the amount of labor per acre on the farms more accurately than the arithmetic mean. The degree to which the line does not describe the relationship is given by the sum of the squares of the deviations from this line. This measure may be written $\Sigma\boxed{2}$. It is obvious that the sum of squares about this line is smaller than that calculated about the line representing the arithmetic mean (compare figures 2 and 3). This indicates that the line in figure 3 is more descriptive of the relationship than the line in figure 2.

The difference between the two sums of squares[4] gives the amount

---

[4] The squared deviations about the arithmetic mean might be averaged and their square root extracted. The result would be the standard deviation. The relationship may be shown algebraically as follows:

$$\sqrt{\frac{\Sigma\boxed{1}}{N}} = \sigma, \text{ more commonly } = \sqrt{\frac{\Sigma y^2}{N}}$$

Likewise, the squared deviations about the line of relationship may be averaged and their square root extracted and expressed algebraically as follows:

$$\sqrt{\frac{\Sigma\boxed{2}}{N}} = S_Y = \sqrt{\frac{\Sigma(y')^2}{N}}$$

This measure, known as the standard error of estimate, is discussed on page 149.

by which the original variability in labor per acre is reduced by considering yields. This difference, $\Sigma\boxed{1} - \Sigma\boxed{2}$, may be expressed as a proportion of the total original variability in labor per acre, $\Sigma\boxed{1}$, and written $\frac{\Sigma\boxed{1} - \Sigma\boxed{2}}{\Sigma\boxed{1}}$. This ratio, measuring the proportion of variability in labor explainable by variations in yield, is commonly represented by r^2 and is known as the coefficient of determination. Its square root, r, is the correlation coefficient. The percentage determination is 100 times r^2.

After the meaning of correlation is understood, the next problem is the examination of various methods of its calculation.

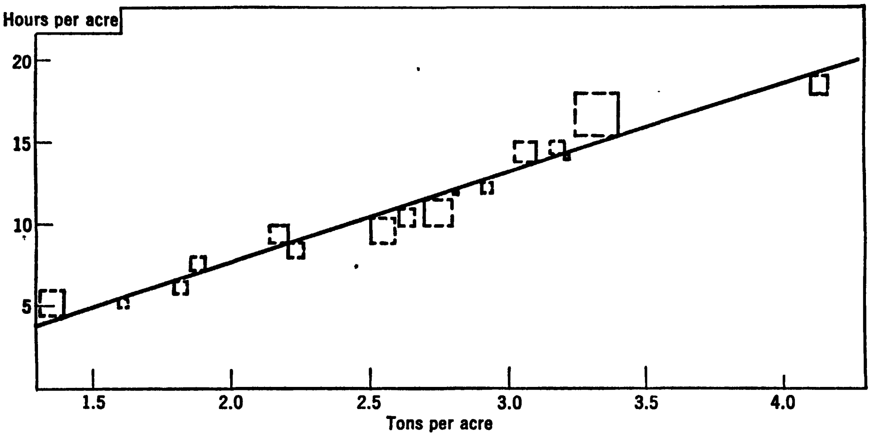


FIGURE 3.—DEVIATIONS AND SQUARES OF DEVIATIONS FROM LINE OF RELATIONSHIP

Yield of Alfalfa and Hours to Harvest an Acre on 16 New York Farms, 1938

The rectangles plotted above and below the line are equilateral. Each side of a square is the amount of labor on a given farm expressed as a deviation from the amount expected from the line of relationship. The area of each rectangle represents the square of this deviation.

LEAST-SQUARES METHOD

Although the least-squares method of obtaining the correlation coefficient is not the easiest or the most commonly used method, it logically follows the above description of the principles involved.

The first step is the calculation of the least-squares line showing the relationship between the two variables, yield and hours of labor. To determine the straight line given by the equation $Y = a + bX$, the values of a and b are calculated by solving two simultaneous equations. Their solution requires the prior calculation of the following quantities: ΣY, ΣX, ΣX^2, and ΣXY, which are given to the left in table 2. The

TABLE 2.—CALCULATION OF THE LEAST-SQUARES LINE OF RELATIONSHIP

YIELD OF ALFALFA AND LABOR PER ACRE ON 16 NEW YORK FARMS, 1938

Determination of the necessary values for

Farm number	Yield, tons per acre X	Labor, hours per acre Y	X^2	XY
1	2.5	9	6.25	22.5
2	2.6	10	6.76	26.0
3	3.2	15	10.24	48.0
4	2.9	12	8.41	34.8
5	1.6	5	2.56	8.0
6	1.4	6	1.96	8.4
7	2.7	10	7.29	27.0
8	1.9	8	3.61	15.2
9	2.2	10	4.84	22.0
10	2.8	12	7.84	33.6
11	3.1	15	9.61	46.5
12	3.4	18	11.56	61.2
13	2.2	8	4.84	17.6
14	1.8	6	3.24	10.8
15	3.2	14	10.24	44.8
16	4.1	18	16.81	73.8
Total	41.6	176	116.06	500.2

Solution of the normal equations

$$Y = a + bX$$

$$Na + b\Sigma X - \Sigma Y = 0$$
$$a\Sigma X + b\Sigma X^2 - \Sigma XY = 0$$

$$16a + 41.6b - 176 = 0$$
$$41.6a + 116.06b - 500.2 = 0$$

Divide by the coefficients of b:

$$\begin{array}{l} 0.384615a + b - 4.230769 = 0 \\ \underline{0.358435a + b - 4.309840 = 0} \\ 0.026180a \qquad + 0.079071 = 0 \end{array}$$

$$a = -3.0203$$

Divide by the coefficients of a:

$$\begin{array}{l} a + 2.600000b - 11.000000 = 0 \\ \underline{a + 2.789904b - 12.024038 = 0} \\ -0.189904b + 1.024038 = 0 \end{array}$$

$$b = 5.3924$$

$$Y = -3.0203 + 5.3924X$$

normal equations and their solution appear to the right of table 2. The line of relationship, or the "regression line,"[5] as it is termed, is given by the equation

$$Y = -3.0203 + 5.3924X$$

and is shown graphically in figure 1.

The second step involves the determination of the estimated hours of labor per acre for each farm, based on the normal relationship shown

---

[5] The term "regression" was introduced by Francis Galton in the latter part of the nineteenth century. Galton correlated the relationship between the heights of fathers and sons. He found that the mean height of the sons of parents of a given type was nearer the mean height of the general population than were their parents' heights. This tendency of the sons to revert back was called regression. The term regression, which originally described biological relationships like the above, through usage gradually came to describe any relationship.

TABLE 3.—CALCULATION OF THE CORRELATION COEFFICIENT BY THE LEAST-SQUARES METHOD

Yield of Alfalfa and Labor per Acre on 16 New York Farms, 1938

Farm number	Calculation of σ_Y^2			Calculation of S_Y^2		
	Labor, hours per acre Y	Deviation from mean y	Deviation from mean squared y^2	Estimated labor, line of relation Y'	Deviation from line y'	Deviation from line squared $(y')^2$
1	9	−2	4	10.5	−1.5	2.25
2	10	−1	1	11.0	−1.0	1.00
3	15	+4	16	14.2	+0.8	0.64
4	12	+1	1	12.6	−0.6	0.36
5	5	−6	36	5.6	−0.6	0.36
6	6	−5	25	4.5	+1.5	2.25
7	10	−1	1	11.5	−1.5	2.25
8	8	−3	9	7.2	+0.8	0.64
9	10	−1	1	8.8	+1.2	1.44
10	12	+1	1	12.1	−0.1	0.01
11	15	+4	16	13.7	+1.3	1.69
12	18	+7	49	15.3	+2.7	7.29
13	8	−3	9	8.8	−0.8	0.64
14	6	−5	25	6.7	−0.7	0.49
15	14	+3	9	14.2	−0.2	0.04
16	18	+7	49	19.1	−1.1	1.21
Total	176	0	252	—	—	22.56

$$\sigma_Y^2 = \frac{252}{16} = 15.7500$$

$$S_Y^2 = \frac{\Sigma(y')^2}{N}$$

$$= \frac{22.56}{16} = 1.4100$$

$$r = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}}$$

$$r = \sqrt{1 - \frac{1.4100}{15.7500}}$$

$$r = \sqrt{1 - 0.0895}$$

$$= \sqrt{0.9105}$$

$$= 0.954$$

by the equation. This set of values is determined by substituting in the equation the yield per acre for each farm, X, and solving for Y, the corresponding estimated normal labor requirements. The values of Y estimated from such an equation are commonly called Y', to distinguish them from the actual values of Y. For farm 1, the yield was 2.5 tons, and the estimated labor required was 10.5 hours [$Y' = -3.0203 + (5.3924)(2.5) = 10.5$]. For farm 6, the estimated labor required was 4.5 hours [$Y' = -3.0203 + (5.3924)(1.4) = 4.5$]. The estimated values[6] for each farm obtained in this way are given to the right of the second double line in table 3.

---

[6] These values might have been read from the straight line in figures 1 or 3.

The next step involves the determination of the differences between the actual hours, Y, and the estimated hours, Y', for each farm. This difference, $Y - Y'$, or y', for farm 1 was -1.5 $(9-10.5 = -1.5)$. The difference[7] for farm 2 was -1.0, and so on. These deviations are then squared, as shown in the last column of table 3. The deviations from the line and their squares are shown in tabular form in the last two columns of table 3; they are shown graphically in figure 3. One side of each of the rectangles represents a deviation; the area of the rectangle, the square of that deviation. The sum of these squared deviations was 22.56; and their average, 1.41, is comparable to the squared standard deviation, σ^2, except that the deviations are taken about the line of relationship rather than the arithmetic mean. This measure is the squared standard error of estimate,[8] or the variance about the line of relationship.

$$S_Y^2 = \frac{\Sigma(y')^2}{N} = \frac{22.56}{16} = 1.41$$

The squared standard deviation or variance about the arithmetic mean, calculated by the usual method, was 15.75 (table 3, left). The deviations and their squares are shown graphically in figure 2. The squared standard deviations about the arithmetic mean, σ_Y^2, and about the "regression" line, S_Y^2, measure the degree to which the mean and the line, respectively, fail to characterize the data. Since $S_Y^2 = 1.41$ and $\sigma_Y^2 = 15.75$, it is clear that the regression line describes the data more accurately than the arithmetic mean. The difference between these two quantities measures the amount of the original variability or variance about the arithmetic mean eliminated by reference to yield per acre. The proportion of this variability eliminated is given by the ratio of this difference to the original variability,

$$\frac{\sigma_Y^2 - S_Y^2}{\sigma_Y^2} = \frac{15.75 - 1.41}{15.75} = 0.91$$

and is the coefficient of determination, r^2. This quantity may also be written:

$$1 - \frac{S_Y^2}{\sigma_Y^2} = 1 - \frac{1.41}{15.75} = 0.91$$

---

[7] The differences are sometimes called "residuals" and described algebraically as $Y - Y'$, y', or z.

[8] Its square root, the standard error of estimate, is the standard deviation about the line of regression

The correlation coefficient is the square root of this quantity:

$$r = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}} = \sqrt{1 - \frac{1.41}{15.75}} = \sqrt{0.91} = \pm 0.95$$

These calculations show that r is either plus or minus 0.95. Whether the sign of the correlation coefficient is plus or minus depends on the line of relationship. Since the line shows that as one variable increases the other also increases, the correlation coefficient reads $r = +0.95$, as above. Had one variable increased as the other decreased, the coefficient would have read $r = -0.95$.

The coefficient of determination, $r^2 = 0.91$, and the correlation coefficient, $r = +0.95$, are individual, abstract numbers, not in terms of tons of hay or hours of labor. The size of these numbers[9] indicates that there was a close relationship between yield and the labor required to harvest the crop. The positive sign of the correlation coefficient indicates that, the greater the yield, the greater the amount of labor required to harvest the crop.[10]

PRODUCT-MOMENT METHOD OF CORRELATION WITH DEVIATIONS FROM ARITHMETIC MEAN

This method is based on relating the deviations of two series from their respective means. This is in distinct contrast to the least-squares method where the amount of correlation was based on the deviations of the dependent variable about the regression line related to deviations

---

[9] When the relationship is positive, the values of r range from 0 to +1.0; and, when negative, from 0 to −1.0. Therefore, the values of r^2 always range from 0 to +1.0. The values of $r^2 = 0.91$ and $r = +0.95$ are relatively high.

[10] The coefficient of determination, which is the ratio of the difference $\sigma_Y^2 - S_Y^2$ to the original variability σ_Y^2, is also given by the ratio of the sums of squared deviations,

$$\frac{\sigma_Y^2 - S_Y^2}{\sigma_Y^2} = \frac{\frac{\Sigma y^2}{N} - \frac{\Sigma(y')^2}{N}}{\frac{\Sigma y^2}{N}} = \frac{\Sigma y^2 - \Sigma(y')^2}{\Sigma y^2}$$

The last expression is the same as $\frac{\Sigma\boxed{1} - \Sigma\boxed{2}}{\Sigma\boxed{1}}$, where $\Sigma\boxed{1}$ is the size of the squared areas based on deviations from the average, Σy^2, and $\Sigma\boxed{2}$ is the size of the squared areas based on deviations from the trend, $\Sigma(y')^2$. The relative size of the sums of these squared areas is shown numerically by 252 and 22.56 (table 3), and graphically, by the rectangles distributed about the average line and about the regression line in figures 2 and 3. From these graphs, one immediately gets the impression that the sizes of the squared areas about the regression line are much smaller than those about the average. This indicates that yield accounts for a considerable amount of the original variability in labor required. When the effect of yield is eliminated, little unaccounted-for variability remains.

about its mean. The coefficients calculated by the two methods are identical.[11] The product-moment method may be written diagrammatically as follows:

$$\text{Correlation coefficient} = \frac{\dfrac{\text{Sum of products of deviations in one variable from its average times the corresponding deviations in the second variable}}{\text{Number of observations}}}{\text{Product of the standard deviations of the two variables}}$$

This may be written algebraically in the following forms:

$$r = \frac{\dfrac{(X_1-AX)(Y_1-AY)+(X_2-AX)(Y_2-AY)+\cdots+(X_N-AX)(Y_N-AY)}{N}}{\sigma_X\sigma_Y}$$

$$r = \frac{\dfrac{\Sigma xy}{N}}{\sigma_X\sigma_Y} = \frac{\Sigma xy}{N\sigma_X\sigma_Y}$$

For the student who is acquainted with the calculation of standard deviations, there is nothing new in the above formula except the so-called product sum, Σxy. This calculation of product sums is relatively simple after one has performed the operations necessary to obtain the two standard deviations.

To determine the standard deviations, it was necessary to get the deviations for all farms from their respective means of yield and hours of labor. The average yield of hay was 2.6 tons; and the average amount of labor to harvest an acre, 11 hours. On farm 1, the yield was 2.5 tons; and the hours to harvest, 9. The deviation for yield was -0.1 $(2.5 - 2.6)$; and for hours, -2 $(9 - 11)$ (table 4). These deviations were squared (0.01 and 4) in the process of obtaining the standard deviations.

The product is obtained by multiplying these two paired deviations together $[(-0.1)(-2) = +0.2]$. For farm 2, the product was 0 $[(0)(-1) = 0]$; and for farm 3, $+2.4$ $[(+0.6)(+4.0) = +2.4]$. The sum of the products for 16 farms was $+42.6$, averaging $+2.6625$, the product moment. The positive signs of the product sum and the product moment indicated that the relationship was positive, that is, that a change in one variable in a given direction was accompanied, on the average, by a change in the second variable in the same direction. For example, when the

---

[11] It can be shown algebraically that the correlation coefficients determined by the least-squares and product-moment methods are identical; therefore,

$$\sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}} = \frac{\Sigma xy}{N\sigma_X\sigma_Y}$$

TABLE 4.—PRODUCT-MOMENT METHOD OF CORRELATION, WITH DEVIATIONS FROM ACTUAL MEAN

YIELD OF ALFALFA AND LABOR PER ACRE ON 16 NEW YORK FARMS, 1938

Farm number	Yield, tons per acre X	Labor, hours per acre Y	Deviations from means		Deviations squared		Product of deviations xy
			x	y	x^2	y^2	
1	2.5	9	−0.1	−2	0.01	4	+ 0.2
2	2.6	10	0	−1	0.00	1	0.0
3	3.2	15	+0.6	+4	0.36	16	+ 2.4
4	2.9	12	+0.3	+1	0.09	1	+ 0.3
5	1.6	5	−1.0	−6	1.00	36	+ 6.0
6	1.4	6	−1.2	−5	1.44	25	+ 6.0
7	2.7	10	+0.1	−1	0.01	1	− 0.1
8	1.9	8	−0.7	−3	0.49	9	+ 2.1
9	2.2	10	−0.4	−1	0.16	1	+ 0.4
10	2.8	12	+0.2	+1	0.04	1	+ 0.2
11	3.1	15	+0.5	+4	0.25	16	+ 2.0
12	3.4	18	+0.8	+7	0.64	49	+ 5.6
13	2.2	8	−0.4	−3	0.16	9	+ 1.2
14	1.8	6	−0.8	−5	0.64	25	+ 4.0
15	3.2	14	+0.6	+3	0.36	9	+ 1.8
16	4.1	18	+1.5	+7	2.25	49	+10.5
Total	41.6	176	0.0	0	7.90	252	42.6
Average	2.6	11	—	—	0.49375	15.75	2.6625

$$\sigma_X = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{7.90}{16}} = \sqrt{0.4937500} = 0.702673$$

$$\sigma_Y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{252}{16}} = \sqrt{15.75000} = 3.96863$$

$$\frac{\Sigma xy}{N} = \frac{42.6}{16} = 2.6625$$

$$r = \frac{\frac{\Sigma xy}{N}}{\sigma_X \sigma_Y} = \frac{2.6625}{0.702673 \times 3.96863} = \frac{2.6625}{2.7886} = +0.955$$

yields were less than average, the labor required to harvest the crop was also less than average; when the yield was above average, the labor required was also above normal.

If the positive deviations in one series are generally accompanied by negative deviations in the other series, the sign of the product sum, product moment, and the correlation coefficient are negative and indicate that, on the average, a change in one variable in a given direction is accompanied by a change in the second variable in the opposite direction.

The "product sum," as this term indicates, is the sum of the products of deviations in terms of tons of hay and hours of labor. This quantity, +42.6, is an expression containing the interrelations between yields per acre and hours of labor. However, this number is meaningless in itself because, although it contains the products of paired variations in which the student is interested, its significance is obscured by the two different units of measurement involved, hours and tons. When the effect of the size of the two units is eliminated, the meaningless product sum, +42.6, is converted into an abstract expression, $r = +0.95$, that has a very definite meaning to the statistician. The conversion of the meaningless product sum to the meaningful abstract correlation coefficient is accomplished by expressing its average, +2.6625, the product moment, as a ratio to the product of the two standard deviations, 0.703 ton and 3.969 hours. The calculation is as follows:

$$r = \frac{\frac{\Sigma xy}{N}}{\sigma_X \sigma_Y} = \frac{+2.6625}{(0.702673)(3.96863)} = +0.955$$

The correlation coefficient, $r = +\ 0.95$, indicates that there was a high degree of relationship between the two series. The sign preceding the correlation coefficient is positive and indicates that, when one variable increased, the other also increased.

Note that with the product-moment method the sign of the correlation coefficient is always correctly indicated by the calculations. In the least-squares method, the sign indicated by calculation is always both plus and minus; and the proper sign has to be determined by inspection of the relationship.

PRODUCT-MOMENT METHOD OF CORRELATION WITHOUT DEVIATIONS

It is not necessary to use deviations from the arithmetic mean to determine the standard deviation.[12] The standard deviations calculated with and without deviations give the same results, because

$$\frac{\Sigma x^2}{N} = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2$$

It is also true that the product sum and product moment can be calculated without first determining the deviations for each series from

---

[12] Page 45.

TABLE 5.—PRODUCT-MOMENT METHOD OF CORRELATION, WITHOUT DEVIATIONS

YIELD OF ALFALFA AND LABOR PER ACRE ON 16 NEW YORK FARMS, 1938

Farm number	Yield, tons per acre X	Labor, hours per acre Y	X^2	Y^2	XY
1	2.5	9	6.25	81	22.5
2	2.6	10	6.76	100	26.0
3	3.2	15	10.24	225	48.0
4	2.9	12	8.41	144	34.8
5	1.6	5	2.56	25	8.0
6	1.4	6	1.96	36	8.4
7	2.7	10	7.29	100	27.0
8	1.9	8	3.61	64	15.2
9	2.2	10	4.84	100	22.0
10	2.8	12	7.84	144	33.6
11	3.1	15	9.61	225	46.5
12	3.4	18	11.56	324	61.2
13	2.2	8	4.84	64	17.6
14	1.8	6	3.24	36	10.8
15	3.2	14	10.24	196	44.8
16	4.1	18	16.81	324	73.8
Total	41.6	176	116.06	2188	500.2
Average	2.6	11	7.25375	136.75	31.2625

Formula 1: Based on sums

$$r=\frac{\Sigma XY-AX\cdot\Sigma Y}{\sqrt{\Sigma X^2-AX\cdot\Sigma X}\sqrt{\Sigma Y^2-AY\cdot\Sigma Y}}$$

$$r=\frac{500.2-2.6\times 176}{\sqrt{116.06-2.6\times 41.6}\sqrt{2188-11\times 176}}$$

$$=\frac{500.2-457.6}{\sqrt{116.06-108.16}\sqrt{2188-1936}}$$

$$=\frac{42.6}{\sqrt{7.9}\sqrt{252}}=\frac{42.6}{2.8107\times 15.875}$$

$$=\frac{42.60}{44.62}=+0.955$$

Formula 2: Based on averages

$$r=\frac{AXY-AX\cdot AY}{\sqrt{AX^2-(AX)^2}\sqrt{AY^2-(AY)^2}}$$

$$=\frac{31.2625-2.6\times 11}{\sqrt{7.25375-(2.6)^2}\sqrt{136.75-(11)^2}}$$

$$=\frac{31.2625-28.6}{\sqrt{7.25375-6.76}\sqrt{136.75-121}}$$

$$=\frac{2.6625}{\sqrt{0.49375}\sqrt{15.75}}$$

$$=\frac{2.6625}{0.70267\times 3.9686}=\frac{2.663}{2.789}$$

$$=+0.955$$

their respective means. The two methods of determining the product moment may be written diagrammatically and algebraically as follows:

$$\text{Product moment}=\frac{\Sigma\left[\begin{pmatrix}\text{Deviations of one}\\ \text{variable from its}\\ \text{arithmetic mean}\end{pmatrix}\begin{pmatrix}\text{Corresponding deviations}\\ \text{of second variable from}\\ \text{its arithmetic mean}\end{pmatrix}\right]}{\text{Number of items}}=\frac{\Sigma xy}{N}$$

$$=\frac{\Sigma\left[\begin{pmatrix}\text{Original}\\ \text{items of}\\ \text{first}\\ \text{variable}\end{pmatrix}\begin{pmatrix}\text{Corresponding}\\ \text{original items}\\ \text{of second}\\ \text{variable}\end{pmatrix}\right]}{\begin{matrix}\text{Number}\\ \text{of items}\end{matrix}}-\begin{pmatrix}\text{Average of first}\\ \text{variable times}\\ \text{average of second}\\ \text{variable}\end{pmatrix}=\frac{\Sigma XY}{N}-AX\cdot AY$$

The quantities required for the calculation of the correlation coefficient are those required to obtain the standard deviations and the product sum. The necessary values, ΣX, ΣY, ΣX^2, ΣY^2, and ΣXY, are given in table 5. The correlation coefficient may be obtained either from these sums or from their corresponding averages, because

$$r=\frac{\Sigma XY-AX\cdot\Sigma Y}{\sqrt{\Sigma X^2-AX\cdot\Sigma X}\sqrt{\Sigma Y^2-AY\cdot\Sigma Y}}=\frac{AXY-AX\cdot AY}{\sqrt{AX^2-(AX)^2}\sqrt{AY^2-(AY)^2}}$$

The student will immediately recognize that the two formulas are identical because the second formula is merely the first with both the numerator and denominator divided by N. The correlation coefficient by both the first and the second formulas was +0.955 (table 5). The coefficients are the same as those obtained by the least-squares[13] method and the product-moment method with deviations.[14]

PRODUCT-MOMENT METHOD WITH DEVIATIONS FROM ASSUMED MEANS OF GROUPED DATA

Prior to the time when calculating and tabulating machines were available, methods were devised to obtain the standard deviations from frequency distributions.[15] In these methods, the deviations were expressed in class intervals. For the calculation of the correlation coefficient from a large number of observations, methods were also developed to obtain the product sum from two interrelated frequency distributions. With these methods of determining standard deviations and the product sum, the calculation of r was shortened considerably.

"Double classification" tables, bivariate frequency distributions, "double-entry" tables, or "correlation" tables were developed to facilitate the calculation of the product sum (table 6). The two series of paired items were grouped into certain class intervals. In the problem under consideration,[16] the yields of alfalfa were grouped in the classes 1.4–1.7 tons, 1.8–2.1 tons, etc.; and the hours of labor, 4–5 hours, 6–7 hours, etc. Consequently, a farm was grouped with consideration to the two factors yield and hours of labor. For the first farm, 9 hours of labor were required to harvest an acre yielding 2.5 tons. The 9 hours would fall in the class 8–9; and the 2.5 tons, in the class 2.2–2.5. In the double-entry table, this farm falls in a compartment with definite, prescribed limits, namely, yield of 2.2–2.5 tons; and labor of 8–9 hours. There were 11 farms falling within these limits. Farm 6, with a yield of 1.4 tons and 6 hours of labor, fell in the compartment with limits of 1.4–1.7 tons and 6–7 hours. The same procedure was followed for each of the 192 farms. This resulted in 15 different frequency distributions, 8 for hours, f_Y; and 7 for yield, f_X (table 6).

The sums of the frequencies added horizontally, which appear in column f_Y, form the usual type of frequency distribution of the 192 farms that one would construct to calculate the average and standard deviation in hours of labor. The arbitrary origin was set at the mid-

---

[13] Table 3, page 148. [14] Table 4, page 152. [15] Page 46.

[16] For illustrative purposes, only 16 farms were included in the previous methods. This method is better illustrated with a larger number of observations; and for this reason, 192 farms were used.

TABLE 6.—PRODUCT-MOMENT METHOD OF CORRELATION, WITH DEVIATIONS FROM ASSUMED MEANS OF GROUPED DATA

YIELD OF ALFALFA AND LABOR PER ACRE, 192 NEW YORK FARMS, 1938

Bold-face numbers are frequencies; italicized numbers are products of deviations and frequencies.

Hours, Y	Tons, X							f_Y	d_Y	f_Yd_Y	$f_Yd_Y^2$	ΣP_{XY}
	1.4–1.7	1.8–2.1	2.2–2.5	2.6–2.9	3.0–3.3	3.4–3.7	3.8–4.1					
18–19					*8* **2**	*16* **2**	*36* **3**	7	+4	+28	112	60
16–17			*−3* **1**	*0* **2**	*27* **9**	*42* **7**	*9* **1**	20	+3	+60	180	75
14–15			*−2* **1**	*0* **10**	*46* **23**	*4* **1**		35	+2	+70	140	48
12–13		*−2* **1**	*−5* **5**	*0* **27**	*12* **12**			45	+1	+45	45	5
10–11		*0* **3**	*0* **17**	*0* **15**	*0* **7**			42	0	0	0	0
8– 9	*3* **1**	*12* **6**	*11* **11**	*0* **5**	*−1* **1**			24	−1	−24	24	25
6– 7	*30* **5**	*12* **3**	*8* **4**	*0* **1**	*−2* **1**			14	−2	−28	56	48
4– 5	*18* **2**	*6* **1**	*6* **2**					5	−3	−15	45	30
f_X	8	14	41	60	55	10	4	192		136	602	291
d_X	−3	−2	−1	0	+1	+2	+3					
f_Xd_X	−24	−28	−41	0	+55	+20	+12	−6				
$f_Xd_X^2$	72	56	41	0	55	40	36	300				
ΣP_{XY}	51	28	15	0	90	62	45	291	Summation, Σ			

$$c_X = \frac{\Sigma f_X d_X}{N} = \frac{-6}{192} = -0.03125$$

$$\sigma_X^2 = \frac{\Sigma f_X d_X^2}{N} - c_X^2 = \frac{300}{192} - (-0.03125)^2$$

$$= 1.5625 - 0.0010 = 1.5615$$

$$\sigma_X = \sqrt{1.5615} = 1.2496$$

$$\frac{\Sigma P_{XY}}{N} = \frac{291}{192} = 1.5156$$

$$r = \frac{\dfrac{\Sigma P_{XY}}{N} - c_X c_Y}{\sigma_X \sigma_Y}$$

$$c_Y = \frac{\Sigma f_Y d_Y}{N} = \frac{136}{192} = 0.7083$$

$$\sigma_Y^2 = \frac{\Sigma f_Y d_Y^2}{N} - c_Y^2 = \frac{602}{192} - (0.7083)^2$$

$$= 3.1354 - 0.5017 = 2.6337$$

$$\sigma_Y = \sqrt{2.6337} = 1.6229$$

$$r = \frac{1.5156 - (-0.03125 \times 0.7083)}{(1.2496)\,(1.6229)}$$

$$= \frac{1.5156 + 0.0221}{2.0280} = \frac{1.5377}{2.0280}$$

$$= +0.76$$

point of the 10–11-hour class. The standard deviation by methods already discussed[17] was

$$\sqrt{\frac{\Sigma f d_Y^2}{N} - \left(\frac{\Sigma f d_Y}{N}\right)^2} = \sqrt{\frac{602}{192} - \left(\frac{136}{192}\right)^2} = 1.62$$

Similarly, the sums of the frequencies added vertically, which appear in line f_X, form the frequency distribution of the 192 farms one would use to obtain the standard deviation of yields. This standard deviation was

$$\sigma_X = \sqrt{\frac{300}{192} - \left(\frac{-6}{192}\right)^2} = 1.25$$

---

[17] Page 48.

These two standard deviations, 1.62 and 1.25, are in terms of their respective class intervals.[18]

The calculation of these standard deviations is already familiar to the student. The difficulty of this method lies in the derivation of the product sum. Confusion arises because of the difficulty of comprehending the triple multiplication of two deviations from arbitrary origins and their corresponding frequency, and also because of the tediousness of the work. In the first row of table 6 are three frequencies, 2, 2, and 3, which deviate +4 class intervals from the arbitrary origin for hours, $d_Y = +4$. Each of these three frequencies was multiplied by this +4 and also by its deviation from the arbitrary origin for yield, +1, +2, and +3, respectively, given in line d_X. The three products were +8, +16, and +36. Their sum, 60, was entered as the first item in the last column to the right, headed product sum, ΣP_{XY}. This is the step that confuses the student. It is difficult to determine which deviation in one variable and which deviation in the other correspond with the particular frequency. In other words, it is visually difficult to follow this system of determining products.

The inability to follow the steps in mechanical procedure often prevents the student from understanding the principles involved. The calculation of the first sum of products, 60, may be shown more clearly as follows:

$$\begin{aligned}\text{Sum of products for first row} &= f(d_X)\,(d_Y) + f(d_X)\,(d_Y) + f(d_X)\,(d_Y)\\ &= 2(+1)\,(+4) + 2(+2)\,(+4) + 3(+3)\,(+4)\\ &= 2(+4) + 2(+8) + 3(+12)\\ &= 8 + 16 + 36\\ &= 60\end{aligned}$$

In the first compartment, the number of farms, f, was 2. The deviation of this group for hours, which was +4 class intervals from the arbitrary origin, is shown in the column of table 6 headed by d_Y. The deviation of this group for yield, which was +1 class interval from the arbitrary origin, is shown directly below this compartment in the row d_X.

In the second compartment, the number of farms was 2. The deviation in the d_Y column was still +4, but in the d_X row, +2.

In the third compartment, containing 3 farms, the deviation in the d_Y column was still +4, but in the d_X row, +3.

The product of the deviations for each farm in the first compartment was +4 [(+1) (+4)]; and the sum of the products for the 2 farms was +8 [2(+4)]. Similarly, the product for the two farms in the second column was 16 [2(+2) (+4) = 16]; and for the three farms in the third compartment, the sum of the products was +36 [3(+3) (+4) = +36]. The sum of products for all 7 farms in these three compartments of the first row was 60 (8 + 16 + 36 = 60). Similarly for the 20 farms in the five compartments in the second row, for which the deviations in hours, d_Y, were +3 class intervals, the sum of products was calculated as follows:

---

[18] The standard deviations in terms of class intervals could be converted to original units by multiplying by their respective class intervals. The standard deviation of labor per acre was 3.24 hours (1.62 × 2 = 3.24); and for yield. 0.50 ton (1.25 × 0.4 = 0.50).

$$\begin{aligned}\text{Sum of products for} &= 1(-1)(+3) + 2(0)(+3) + 9(+1)(+3) + 7(+2)(+3) + 1(+3)(+3)\\ &= -3 + 0 + 27 + 42 + 9\\ \text{second row} &= 75\end{aligned}$$

This sum, 75, is entered in the column headed ΣP_{XY} at the right side of table 6. The sums of the products for the farms in each line of the table were calculated in a similar manner. The total of the sums of products for each row was the product sum for the 192 farms. It was 291 in terms of class intervals.

The product sum, ΣP_{XY}, may be obtained from the sums of products of the columns instead of the rows. The calculation for the first column was as follows:

$$\begin{aligned}\text{Sum of products} &= f(d_Y)(d_X) + f(d_Y)(d_X) + f(d_Y)(d_X)\\ \text{for first column} &= 1(-1)(-3) + 5(-2)(-3) + 2(-3)(-3)\\ &= 1(+3) + 5(+6) + 2(+9)\\ &= 51\end{aligned}$$

The same procedure is followed for the remaining six columns.The product sum for all farms was 291, the same as that obtained from the rows. Because of visual difficulty involved in its computation, it is generally advisable to calculate the product sum from both the rows and the columns. If the results are not identical, an error has been made.

The calculation of the product moment used in obtaining the correlation coefficient consists of dividing this product sum by the number of farms and correcting for the use of arbitrary origins. The correction for a squared standard deviation is the square of the quantity $\Sigma fd/N$. For the product moment, the correction is $\left(\frac{\Sigma fd_X}{N}\right)\left(\frac{\Sigma fd_Y}{N}\right)$. Therefore, the corrected product moment was:

$$\begin{aligned}p_{XY} &= \frac{\Sigma P_{XY}}{N} - \left(\frac{\Sigma fd_X}{N}\right)\left(\frac{\Sigma fd_Y}{N}\right)\\ &= \frac{291}{192} - \left(\frac{136}{192} \times \frac{-6}{192}\right)\\ &= 1.5156 - (0.7083 \times -0.03125)\\ &= 1.5377\end{aligned}$$

The correlation coefficient is merely the ratio of this product moment to the product of the standard deviations:

$$r = \frac{1.5377}{1.2496 \times 1.6229}$$
$$= 0.76$$

The product sum and standard deviations were in terms of class intervals. The conversion of these values to their original units of tons and hours would not have changed the size of the correlation coefficient. The class intervals being the same in both the numerator and denominator, they cancel out as follows:

$$r = \frac{1.5377\left(\cancel{\begin{matrix}\text{Class interval}\\ \text{in tons}\end{matrix}}\right)\left(\cancel{\begin{matrix}\text{Class interval}\\ \text{in hours}\end{matrix}}\right)}{1.25\left(\cancel{\begin{matrix}\text{Class interval}\\ \text{in tons}\end{matrix}}\right)1.62\left(\cancel{\begin{matrix}\text{Class interval}\\ \text{in hours}\end{matrix}}\right)}$$

METHODS COMMONLY USED

In practice, the size of the series and the availability of mechanical equipment affect the choice of the method of calculation. When the number of observations is large and mechanical equipment is not available, the data are usually grouped, and the product-moment method with deviations from assumed means is employed. The double-entry table has another decided advantage over other methods. The arrangement of the frequencies in the table resembles a scatter diagram and shows immediately whether the relationship is positive or negative and whether linear or non-linear. This question of linearity is a very important problem which confronts the student as he progresses farther in his study of relationships.

Small series of data are not grouped. The product-moment method *with* deviations from arithmetic means is most frequently used when mechanical equipment is not available, but probably requires as much time as the product-moment method *without* deviations, if not more. The latter method has the advantage that the confusion due to various combinations of plus and minus signs is eliminated. The use of deviations has some advantage when the size of one or both variables is large, because it is somewhat easier to square the deviation of a large number than to square the large number itself.

If mechanical equipment is available, the product-moment method *without* deviations is superior to other methods, because the sums of squares and the sums of products of the original numbers can be obtained accurately and easily from the equipment. This advantage is about the same whether the size of variables is large or small. When the correlation involves several hundred or more paired observations, this method with the use of tabulating equipment has an overwhelming advantage over all others.

The least-squares method is little used in practice. Its value is chiefly to illustrate the principle involved in correlation.

COEFFICIENTS OF NON-DETERMINATION AND ALIENATION

The coefficient of determination, r^2, is the proportion of the total variation or variance in hours of labor accounted for by differences in yield. If the total variability is considered to be 1 and the coefficient of determination 0.91, the difference, 0.09, is a measure of the amount of variability unaccounted for by yield.[19] This measure is known as the coefficient of non-determination or unaccounted-for variability, k^2. The square root of the coefficient of non-determination is termed the coefficient of alienation, k. These relationships may be written algebraically as follows:

Coefficient of correlation[20] $= r = 0.955$ (table 5)
Coefficient of determination $= r^2 = 0.912$
Coefficient of non-determination $= k^2 = 1 - r^2 = 1 - 0.912 = 0.088$
Coefficient of alienation $= k = \sqrt{k^2} = \sqrt{0.088} = 0.297$

REGRESSION

In the least-squares method of calculating the correlation coefficient, it was necessary first to determine the straight line

$$Y = -3.0203 + 5.3924X \text{ (table 2)}$$

This equation, commonly known as the regression equation, describes the average number of hours of labor required to harvest a crop with a given yield. If the two standard deviations and correlation coefficients are available, it is not necessary to follow the procedure used in table 2 to determine the regression equation. With these three values, the regression equation may be determined from the following expression:

$$\begin{aligned} Y - AY &= r\left(\frac{\sigma_Y}{\sigma_X}\right)(X - AX) \\ Y - 11 &= 0.955\left(\frac{3.969}{0.703}\right)(X - 2.6) \\ Y - 11 &= (5.39)(X - 2.6) \\ Y &= 5.39X - 14.0 + 11.0 \\ Y &= -3.0 + 5.39X \end{aligned}$$

This equation is the same as that given above and in table 2.

---

[19] This might be due to type of harvesting machinery used, weather, or other factors.

[20] This coefficient of correlation is sometimes called "simple," "total," or "gross correlation."

This equation has two numerical values: the constant, −3.0; and the regression coefficient, +5.39. The regression coefficient is the most important term and has the same sign as the correlation coefficient. It measures the amount of change in hours of labor required to harvest an acre for each additional ton of yield. When the yield increases 1 ton per acre, the labor required in harvesting an acre increases 5.39 hours. The constant −3.0 merely determines the position of the line on the vertical scale.

This regression coefficient can also be obtained from product sums[21] and sums of squares of deviations given in table 4, as follows:

$$b_{YX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{42.6}{7.90} = 5.39$$

The common symbol for the regression coefficient is b, and the order of the subscripts Y and X indicates that b is the rate of change in Y in terms of X. The coefficient b_{XY} would be the reverse of that above; that is, b_{XY} is the rate of change in X in terms of Y.

ADVANTAGES AND DISADVANTAGES

The correlation method is merely an averaging process by which an average relationship is measured. It has an advantage that it is adapted to small amounts of data. The correlation coefficient summarizes the *degree* of relationship in one number. The regression coefficient summarizes the *nature* of the relationship in one number. Methods of testing reliability of the two coefficients are relatively easy.

The correlation method has the disadvantage that it always assumes linear relationship regardless of whether that assumption is correct. The coefficients are difficult to calculate. The results of correlation analysis are difficult to understand. They are often misinterpreted. Tabular presentation of relationships is usually more effective than the use of correlation and regression coefficients, even when the latter are thoroughly understood.

USES

The primary use of a correlation coefficient is to show with one number the degree of relationship between two variables. These coefficients

---

[21] It can be demonstrated that regression coefficients determined by the various methods above will always be identical.

$$b_{YX} = r\left(\frac{\sigma_Y}{\sigma_X}\right) = \left(\frac{\frac{\Sigma xy}{N}}{\sigma_X\sigma_Y}\right)\left(\frac{\sigma_Y}{\sigma_X}\right) = \frac{\frac{\Sigma xy}{N}}{\sigma_X^2} = \frac{\frac{\Sigma xy}{N}}{\frac{\Sigma x^2}{N}} = \frac{\Sigma xy}{\Sigma x^2}$$

range from +1.0 to 0 to −1.0. When the coefficient is +1 or −1, there is perfect positive or negative relationship between the two series.

Reed correlated July rainfall with the yield of corn in Ohio for the period 1854–1913. The coefficient of correlation, $r = +0.53$, indicated that, in years of heavy rainfall, corn yield tended to be high (table 7). Conversely, in dry years, yields were low. The correlation was not high because characteristic conditions of other months and other factors also influenced the yield.

TABLE 7.—EXAMPLES OF CORRELATION COEFFICIENTS USED IN VARIOUS STUDIES

Variables associated	Correlation coefficient
July rainfall and yield of corn in Ohio, 1854–1913*	+0.53
Circumference of trunk of peach trees and weight of top†	+0.92
Weight and length of ears of leaming corn‡	+0.87
Land values, 1920, and percentage of all farm land in corn§	+0.87
Land values, 1920, and percentage of all farm land in pasture§	−0.75
Hog prices and hog receipts, Chicago‖	−0.40
Yield of wheat and hours of labor¶	+0.04

---

* Reed, W. G., The Coefficient of Correlation, American Statistical Association, Vol. 15, New Series, No. 117, p. 674, June 1917.

† Tufts, W. P., Pruning Young Deciduous Fruit Trees, California Agricultural Experiment Station Bulletin 313, p. 116, October 1919.

‡ Davenport, E., Principles of Breeding, p. 461, 1907.

§ Sarle, C. F., Comparative Study of Farm Land Value in Iowa, unpublished manuscript, p. 15, August 1924.

‖ Wallace, H. A., Agricultural Prices, p. 93, 1920.

¶ Tolley, H. R., Black, J. D., Ezekiel, M. J. B., Input as Related to Output in Farm Organization and Cost-of-Production Studies, United States Department of Agriculture, Department Bulletin No. 1277, p. 23, September 18, 1924.

Tufts found that in California the weight of the top of peach trees was correlated with the circumference of the trunk, $r = +0.92$. The coefficient was positive and very high, indicating that differences in size from tree to tree were uniform for the different parts of the tree. Other examples of correlation coefficients appear in table 7.

The correlation coefficients summarize in one number important relationships between two variables. The usefulness of these coefficients depends in part on a wide knowledge of the meaning of this "yardstick," together with its limitations.

The coefficient of determination, r^2, has all the limitations of the correlation coefficient, r, but has one distinct advantage. A measure of the proportion of the variability in one thing explainable by another is more easily understood than the square root of this ratio.

TABLE 8.—COMPARISON OF COEFFICIENTS OF CORRELATION, DETERMINATION, AND NON-DETERMINATION

Correlation coefficient r	Coefficient of determination r^2	Coefficient of non-determination $(1 - r^2)$
0.10	0.01	0.99
0.20	0.04	0.96
0.50	0.25	0.75
0.80	0.64	0.36
0.90	0.81	0.19
0.95	0.90	0.10

Laymen erroneously interpret correlation coefficients as percentages of determination. For example, a correlation coefficient of 0.50 is assumed to explain one-half the variability, though actually only one-fourth of the variability is accounted for (table 8). Therefore, other factors account for three-fourths of the variability in the original series. A correlation coefficient of 0.50 is not nearly so significant as most persons assume it to be. A correlation coefficient of 0.90 indicates that 81 per cent of the variability has been accounted for, and only 19 per cent remains unaccounted for. A correlation coefficient of 0.20 indicates that 96 per cent of the variations are due to factors other than that considered. Probably fewer errors in conclusions and generalizations would arise if the percentage determination were used more and correlation coefficients less.

The primary use of the regression equation is to describe the nature of the relationships and to show the rates of change in one factor in terms of another. Ogburn studied the cost of living in 1916 and found that the equation for the relationship between income and savings was as follows:

$$\text{Deficit or surplus} = -166.45 + 0.144\ (\text{annual income})$$
$$Y = -166.45 + 0.144X$$

The nature of this relationship, indicated by the equation, was that, with increasing income, the amount saved increased. The equation also describes the rate of this change, that is, the amount saved for each

TABLE 9.—REGRESSION EQUATIONS

RELATION OF INCOME AND SIZE OF FAMILY TO OTHER FACTORS*

Dependent variable Y	Independent variable X	Equation $Y=a+bX$
Deficit or surplus, dollars	Annual family income, dollars	$Y= -166.45+0.144X$
Food cost per adult per day, dollars	Annual family income, dollars	$Y= 0.23+0.00014X$
Expenditures for:		
Food, per cent of total	Annual family income, dollars	$Y= 53.08-0.0113X$
Rent, per cent of total	Annual family income, dollars	$Y= 21.14-0.0013X$
Fuel and light, per cent of total	Annual family income, dollars	$Y= 7.19-0.0013X$
Family clothing, per cent of total	Annual family income, dollars	$Y= 6.70+0.0035X$
Deficit or surplus, dollars	Size of family, number of persons	$Y= 81.09-24.22X$
Food cost per adult per day, dollars	Size of family, number of persons	$Y= 0.62-0.069X$
Expenditures for:		
Food, per cent of total	Size of family, number of persons	$Y= 34.31+1.71X$
Rent, per cent of total	Size of family, number of persons	$Y= 20.66-0.297X$
Fuel and light, per cent of total	Size of family, number of persons	$Y= 5.52+0.05X$
Family clothing, per cent of total	Size of family, number of persons	$Y= 9.03+0.496X$

---

* Ogburn, W. F., Analysis of the Standard of Living in the District of Columbia in 1916, Quarterly Publications of the American Statistical Association, New Series, No. 126, Vol. XVI, pp 374–389, June 1919.

additional dollar of income, 14 cents. The series of regression equations given in table 9 are not, in themselves, very informative to most readers, including statisticians. However, it is possible to translate the relatively unintelligible equation into a simple form that is intelligible to most persons. For instance, from the equation for savings given above, the amount of surplus or deficit from an income of \$1,000 can be determined by substituting \$1,000 in the equation as the value of X. This amount was −\$22.45.

$$\begin{aligned} Y &= -166.45 + 0.144 \times 1{,}000 \\ &= -166.45 + 144.00 \\ &= -22.45 \end{aligned}$$

TABLE 10.—TABULAR PRESENTATION OF RELATIONSHIPS DESCRIBED BY REGRESSION EQUATIONS IN TABLE 9

Family income, dollars	Deficit or surplus, dollars	Daily food cost, cents per adult	Percentage of expenditures for			
			Food	Rent	Fuel and light	Family clothing
\$1,000	\$ −22.45	37.0	41.8	19.8	5.9	10.2
1,250	13.55	40.5	39.0	19.5	5.6	11.1
1,500	49.55	44.0	36.1	19.2	5.2	12.0

This indicates that families with $1,000 incomes spent $22.45 more than they received. Presumably, they purchased goods on credit that were paid for out of next year's income, or not at all. Families with $1,250 incomes saved $13.55; and those with $1,500 incomes, $49.55.

A table of values worked from the regression equation probably shows the relationship more clearly than the equation itself. These values were calculated showing savings, cost of food, and distribution of expenditures for families with three different incomes (table 10).

The cost of food per person per day increased from 37 to 44 cents as the income increased from $1,000 to $1,500 (table 10). The proportion of the total expenditures for food decreased from 42 to 36 per cent as income increased.

Regression equations between size of the family and various factors are also given in table 9. These relationships could likewise be shown in a table similar to table 10.

Regression equations and tables derived from them are frequently more valuable but are less widely used than the correlation coefficients.

CHAPTER 10

MULTIPLE CORRELATION

The correlation and regression coefficients examined in the last chapter measured the degree and nature of the effect of one variable on another. While it is useful to know how some phenomenon is influenced by another, it is also important to know how this phenomenon is affected by several other variables. In nature, relationships tend to be complex rather than simple. One variable is related to a great number of others, many of which may be interrelated among themselves. For example, the growth of vegetation is related to temperature and rainfall which, in turn, may be related to each other. Certain kinds of wild game occur in greatest numbers in areas of plentiful food supply and heavy rainfall. The greatest food supply, in general, is in the areas of heaviest rainfall. The occurrence of game may also be related to temperature, amount of cover, and other factors which may or may not be interrelated. Whether phenomena be biological, physical, chemical, or economic, they are affected by a multiplicity of causal factors. It is part of the statistician's task to determine the effect of one cause, of two or more causes acting separately or simultaneously, or of one cause when the effect of others is eliminated. Multiple ~~correlation~~ analysis studies the effect of two or more factors which may or may not be interrelated, but the effects of which are separate and distinct.

MEANING OF MULTIPLE CORRELATION

The simple correlation coefficient, r, compares the variability about a fitted straight line to the variability about the arithmetic average as measured by the standard deviation.[1] The multiple correlation coefficient, R, compares the variability about a fitted plane, solid, or hyperplane to variability about the arithmetic average as measured by the standard deviation.

The two types of coefficients may be described diagrammatically as follows:

---

[1] Pages 143 to 149.

$$\text{Simple correlation coefficient} = \sqrt{1 - \frac{\text{Sum of squares of deviations about the best-fitting straight line}}{\text{Sum of squares of deviations about the arithmetic mean}}}$$

$$\text{Multiple correlation coefficient} = \sqrt{1 - \frac{\text{Sum of squares of deviations about the best-fitting plane, solid, or hyperplane}}{\text{Sum of squares of deviations about the arithmetic mean}}}$$

The relationships may be expressed algebraically as follows:

$$r_{12} = \sqrt{1 - \frac{S^2_{1.2}}{\sigma^2_1}}, \qquad R_{1.23} = \sqrt{1 - \frac{S^2_{1.23}}{\sigma^2_1}}, \qquad R_{1.234} = \sqrt{1 - \frac{S^2_{1.234}}{\sigma^2_1}}$$

The three expressions are identical except for $S^2_{1.2}$, $S^2_{1.23}$, and $S^2_{1.234}$, the so-called standard errors of estimate.[2] The first, $S^2_{1.2}$, represents the average of the squares of the deviations about the straight line $X_1 = a + b_{12}X_2$. The second, $S^2_{1.23}$, represents the average of the squares about the plane $X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$; and the third, $S^2_{1.234}$, about the solid $X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$.

For two variables, the relationship of the independent variable,[3] X_2, to the dependent variable, X_1, is approximated by the first equation; and $S^2_{1.2}$ is a measure of the degree to which the straight line does not describe this relationship. Likewise, the relationship of the two independent variables, X_2 and X_3, to the dependent variable, X_1, is approximated by the second equation; and $S^2_{1.23}$ is a measure of the degree to which the plane[4] does not describe this relationship. The relationship

---

[2] The subscript 1.2 to the standard error of estimate, $S_{1.2}$, indicates that $S_{1.2}$ is the amount of variability in X_1 about the line of relationship between X_2 and X_1. Interpreted another way, $S_{1.2}$ measures the variability in X_1, with the effect of X_2 eliminated.

Similarly, $S_{1.23}$ is the amount of variability in X_1 about the "plane" of relationship between X_2, X_3, and X_1. More simply, $S_{1.23}$ measures the variability in X_1, with the effects of X_2 and X_3 eliminated.

Likewise, $S_{1.234}$ measures the variability in X_1, with the effects of X_2, X_3, and X_4 eliminated.

[3] The terms independent and dependent variables are arbitrarily applied to the factors that are to be associated; for instance, if one were studying size of farms and income, the size of farms is generally considered the independent variable; and income, the dependent variable. That is, variations in income depend on variations in size.

[4] With one independent variable, the relationship is described algebraically as follows: $X_1 = a + b_{12}X_2$; and described geometrically by a straight line.

With two independent variables, the relationship is described algebraically by the equation $X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$; and described geometrically by a plane surface.

With three independent variables, the relationship is described algebraically by of the three independent variables, X_2, X_3, and X_4, to the dependent variable, X_1, is approximated by the third equation; and $S^2_{1.234}$ is a measure of the degree to which the solid does not describe this relationship. In each case, the squared standard errors of estimate, $S^2_{1.2}$, $S^2_{1.23}$, and $S^2_{1.234}$, measure the unaccounted-for squared variability.

In both simple and multiple correlation coefficients, the unaccounted-for squared variability is expressed as a proportion of the total squared variability about the average, σ_1^2, and subtracted from 1 to obtain the accounted-for squared variability designated as r^2 and R^2.

THE DETERMINATION OF *R*

The only new problem in calculating the multiple correlation coefficient is to determine the value of $S^2_{1.23}$ or $S^2_{1.234}$. It can be demonstrated that $S^2_{1.234} = \sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}$, and diagrammatically that

$$\begin{bmatrix}\text{Standard}\\\text{error of}\\\text{estimate}\\\text{dependent}\\\text{variable}\end{bmatrix}^2 = \begin{bmatrix}\text{Standard}\\\text{deviation}\\\text{dependent}\\\text{variable}\end{bmatrix}^2 - \begin{bmatrix}\text{Regression}\\\text{coefficient}\\\text{dependent}\\\text{variable in}\\\text{terms of}\\\text{first}\\\text{independent}\\\text{variable}\end{bmatrix}\begin{bmatrix}\text{Product}\\\text{moment,}\\\text{dependent}\\\text{and first}\\\text{independent}\\\text{variable}\end{bmatrix}$$

$$- \begin{bmatrix}\text{Regression}\\\text{coefficient,}\\\text{dependent}\\\text{variable in}\\\text{terms of}\\\text{second}\\\text{independent}\\\text{variable}\end{bmatrix}\begin{bmatrix}\text{Product}\\\text{moment,}\\\text{dependent}\\\text{and second}\\\text{independent}\\\text{variable}\end{bmatrix} - \begin{bmatrix}\text{Regression}\\\text{coefficient,}\\\text{dependent}\\\text{variable in}\\\text{terms of}\\\text{third}\\\text{independent}\\\text{variable}\end{bmatrix}\begin{bmatrix}\text{Product}\\\text{moment,}\\\text{dependent}\\\text{and third}\\\text{independent}\\\text{variable}\end{bmatrix}$$

The formula for the multiple correlation coefficient,

$$R^2_{1.234} = 1 - \frac{S^2_{1.234}}{\sigma_1^2} \quad \text{or} \quad \frac{\sigma_1^2 - S^2_{1.234}}{\sigma_1^2}$$

may also be written

$$R^2_{1.234} = \frac{b_{12.34}p_{12} + b_{13.24}p_{13} + b_{14.23}p_{14}}{\sigma_1^2}$$

The only new expressions are partial regression coefficients,[5] $b_{12.34}$, $b_{13.24}$,

---

the equation $X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$; and geometrically by an unbounded solid. This solid, which theoretically would exist in four-dimensional space, possesses the linear characteristics of the straight line and the plane, but is beyond the imagination of most persons.

[5] The subscript 12.34 to the partial regression coefficient $b_{12.34}$ indicates that $b_{12.34}$ measures the rate of change in X_1 with a unit change in X_2, with the effects of X_3 and X_4 eliminated. The first digit of the subscript always indicates the dependent variable. The second digit refers to the independent variable whose effects are measured by the coefficient. The digits to the right of the decimal refer to those independent variables whose effects are eliminated or held constant.

and $b_{14.23}$. The product moment[6] p_{12} is the average of the products of the deviations from their means of the dependent variable X_1, and the first independent variable X_2. The product moment p_{12} is given by the expressions[7]

$$p_{12} = \frac{\Sigma x_1 x_2}{N} = \left[\frac{\Sigma X_1 X_2}{N} - \left(\frac{\Sigma X_1}{N}\right)\left(\frac{\Sigma X_2}{N}\right)\right] = AX_1X_2 - AX_1AX_2$$

The product moments p_{13} and p_{14} are obtained in a similar manner.

The regression coefficients $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$ are determined from the solution of the following three simultaneous equations:

$$\begin{aligned} \sigma_2^2 b_{12.34} + p_{23} b_{13.24} + p_{24} b_{14.23} &= p_{12} \\ p_{23} b_{12.34} + \sigma_3^2 b_{13.24} + p_{34} b_{14.23} &= p_{13} \\ p_{24} b_{12.34} + p_{34} b_{13.24} + \sigma_4^2 b_{14.23} &= p_{14} \end{aligned}$$

It will be noted that these three equations include various product moments and standard deviations which can be calculated from the four variables. With these values known, the equations can be solved for the values of the three regression coefficients, $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$. When these regression coefficients become known, they may be substituted in the formula for the multiple correlation coefficient on page 168. This is a relatively simple calculation.

The calculation of the regression coefficients is somewhat laborious. This task logically divides itself into three parts: (*a*) the determination of the products and the squares of the variables; (*b*) the calculation of product moments and squared standard deviations; and (*c*) the solution of the simultaneous equations.

In studying the factors[8] affecting the price of No. 1 Northern spring wheat at Minneapolis—the dependent variable, X_1—three independent variables were used. These were: the price of imported wheat at Liverpool, England, X_2; the United States production of wheat, X_3; and world wheat production, X_4. The prices and productions and their first differences are given in table 1.

PRODUCTS AND SQUARES

It is necessary to determine the products, X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, and X_3X_4; and the squares, X_1^2, X_2^2, X_3^2, and X_4^2. These are arranged in an orderly manner in table 2. The first four columns are the squares of the first differences of prices and production shown in table 1;

---

[6] The product moment is the same as that described on page 151.

[7] Page 154.

[8] Trends were eliminated by using first differences. The first differences were the independent and dependent variables X_1, X_2, X_3, and X_4.

and the last six columns are all the possible products of these first differences.

For instance, in the crop year 1891–1892, the Minneapolis price of

TABLE 1.—MINNEAPOLIS AND LIVERPOOL PRICES AND UNITED STATES AND WORLD PRODUCTION OF WHEAT, AND THEIR FIRST DIFFERENCES, 1892–1913

Crop year	Crop year prices,* cents per bushel		Production,† 0,000,000 bushels		First differences			
					Prices		Production	
	No. 1 Northern Spring at Minneapolis	Spot, red at Liverpool	United States	World, including Russia	Minneapolis X_1	Liverpool X_2	United States X_3	World and Russia X_4
1891–92	84	114	68	233				
1892–93	66	85	61	248	− 18	− 29	− 7	+ 15
1893–94	59	73	51	256	− 7	− 12	− 10	+ 8
1894–95	61	70	54	259	+ 2	− 3	+ 3	+ 3
1895–96	57	78	54	248	− 4	+ 8	0	− 11
1896–97	71	89	52	254	+ 14	+ 11	− 2	+ 6
1897–98	94	117	61	231	+ 23	+ 28	+ 9	− 23
1898–99	69	85	77	309	− 25	− 32	+ 16	+ 78
1899–00	68	87	66	287	− 1	+ 2	− 11	− 22
1900–01	73	86	60	272	+ 5	− 1	− 6	− 15
1901–02	72	88	76	294	− 1	+ 2	+ 16	+ 22
1902–03	76	89	69	314	+ 4	+ 1	− 7	+ 20
1903–04	90	90	66	336	+ 14	+ 1	− 3	+ 22
1904–05	111	97	56	320	+ 21	+ 7	− 10	− 16
1905–06	84	98	71	339	− 27	+ 1	+ 15	+ 19
1906–07	84	94	74	357	0	− 4	+ 3	+ 18
1907–08	107	110	63	327	+ 23	+ 16	− 11	− 30
1908–09	116	122	64	325	+ 9	+ 12	+ 1	− 2
1909–10	108	117	68	371	− 8	− 5	+ 4	+ 46
1910–11	103	107	63	365	− 5	− 10	− 5	− 6
1911–12	108	114	62	365	+ 5	+ 7	− 1	0
1912–13	86	112	73	394	− 22	− 2	+ 11	+ 29
1913–14	88	106	75	416	+ 2	− 6	+ 2	+ 22
Total	—	—	—	—	+ 4	− 8	+ 7	+ 183
Average	—	—	—	—	+ 0.18	− 0.36	+ 0.32	+ 8.32

* Timoshenko, V. P., Wheat Prices and the World Wheat Market, Cornell University Agricultural Experiment Station, Memoir 118, pp. 98–99, December 1928.

† Agricultural Statistics 1939, pp. 9, 15.

wheat was 84 cents per bushel, and the following year, 66 cents; the difference was −18 cents (table 1). The square of the first difference was 324 (column 1, table 2). The squares were similarly computed for X_2, X_3, and X_4 for each of the 22 years.

The first differences for the Minneapolis and Liverpool prices, −18 and −29, were multiplied to obtain the product 522, X_1X_2 (column 5, table 2). Similar computations were made for all the products and all the years.

TABLE 2.—PRODUCTS AND SQUARES* REQUIRED FOR CALCULATING MULTIPLE CORRELATION

First Differences in the Price and Production of Wheat†

Crop year	Squares				Products					
	X_1^2	X_2^2	X_3^2	X_4^2	X_1X_2	X_1X_3	X_1X_4	X_2X_3	X_2X_4	X_3X_4
1892–93	324	841	49	225	522	126	− 270	203	− 435	− 105
1893–94	49	144	100	64	84	70	− 56	120	− 96	− 80
1894–95	4	9	9	9	− 6	6	6	− 9	− 9	9
1895–96	16	64	0	121	− 32	0	44	0	− 88	0
1896–97	196	121	4	36	154	− 28	84	− 22	66	− 12
1897–98	529	784	81	529	644	207	− 529	252	− 644	− 207
1898–99	625	1,024	256	6,084	800	− 400	− 1,950	− 512	− 2,496	1,248
1899–00	1	4	121	484	− 2	11	22	− 22	− 44	242
1900–01	25	1	36	225	− 5	− 30	− 75	6	15	90
1901–02	1	4	256	484	− 2	− 16	− 22	32	44	352
1902–03	16	1	49	400	4	− 28	80	− 7	20	− 140
1903–04	196	1	9	484	14	− 42	308	− 3	22	− 66
1904–05	441	49	100	256	147	− 210	− 336	− 70	− 112	160
1905–06	729	1	225	361	− 27	− 405	− 513	15	19	285
1906–07	0	16	9	324	0	0	0	− 12	− 72	54
1907–08	529	256	121	900	368	− 253	− 690	− 176	− 480	330
1908–09	81	144	1	4	108	9	− 18	12	− 24	− 2
1909–10	64	25	16	2,116	40	− 32	− 368	− 20	− 230	184
1910–11	25	100	25	36	50	25	30	50	60	30
1911–12	25	49	1	0	35	− 5	0	− 7	0	0
1912–13	484	4	121	841	44	− 242	− 638	− 22	− 58	319
1913–14	4	36	4	484	− 12	4	44	− 12	− 132	44
Total	4,364	3,678	1,593	14467	+2,928	− 1,233	− 4,847	− 204	− 4,674	+2,735
Average	198.3636	167.1818	72.4091	657.5909	133.0909	− 56.0455	− 220.3182	− 9.2727	− 212.4545	124.3182

---

* A method of calculating sums of products and sums of squares with tabulating equipment is given on page 425, Appendix B.

† Table 1.

The next step was the addition of these squares and products[9] and the calculation of their averages. For example, the sum of the squares of the dependent variable X_1 was $\Sigma X_1^2 = 4{,}364$, which, divided by 22, gave an average of squares $AX_1^2 = 198.3636$. Similarly, the sum of the products of X_1 and X_2 was $\Sigma X_1X_2 = +2{,}928$, and the average was $AX_1X_2 = +133.0909$ (table 2).

TABLE 3.—CALCULATION OF THE 6 PRODUCT MOMENTS AND OF THE 4 SQUARED STANDARD DEVIATIONS

FIRST DIFFERENCE OF PRICE AND PRODUCTION OF WHEAT*

Product moments

$$p_{12} = AX_1X_2 - (AX_1 \cdot AX_2) = 133.0909 - (0.1818 \times -0.3636) = 133.0909 + 0.0661 = 133.1570$$
$$p_{13} = AX_1X_3 - (AX_1 \cdot AX_3) = -56.0455 - (0.1818 \times 0.3182) = -56.0455 - 0.0578 = -56.1033$$
$$p_{14} = AX_1X_4 - (AX_1 \cdot AX_4) = -220.3182 - (0.1818 \times 8.3182) = -220.3182 - 1.5122 = -221.8304$$
$$p_{23} = AX_2X_3 - (AX_2 \cdot AX_3) = -9.2727 - (-0.3636 \times 0.3182) = -9.2727 + 0.1157 = -9.1570$$
$$p_{24} = AX_2X_4 - (AX_2 \cdot AX_4) = -212.4545 - (-0.3636 \times 8.3182) = -212.4545 + 3.0245 = -209.4300$$
$$p_{34} = AX_3X_4 - (AX_3 \cdot AX_4) = 124.3182 - (0.3182 \times 8.3182) = 124.3182 - 2.6469 = 121.6713$$

Squared standard deviations

$$\sigma_1^2 = AX_1^2 - (AX_1)^2 = 198.3636 - (0.1818)^2 = 198.3636 - 0.0331 = 198.3305$$
$$\sigma_2^2 = AX_2^2 - (AX_2)^2 = 167.1818 - (-0.3636)^2 = 167.1818 - 0.1322 = 167.0496$$
$$\sigma_3^2 = AX_3^2 - (AX_3)^2 = 72.4091 - (0.3182)^2 = 72.4091 - 0.1013 = 72.3078$$
$$\sigma_4^2 = AX_4^2 - (AX_4)^2 = 657.5909 - (8.3182)^2 = 657.5909 - 69.1925 = 588.3984$$

* Table 2.

PRODUCT MOMENTS AND SQUARED STANDARD DEVIATIONS

The product moment p_{12} was determined by the following formula:

$$p_{12} = AX_1X_2 - AX_1AX_2$$

The calculations were as follows:

$$p_{12} = \underset{\substack{\text{(table 2,}\\ \text{column 5)}}}{133.0909} - \underset{\substack{\text{(table 1,}\\ \text{column 5)}}}{(+0.1818)}\underset{\substack{\text{(table 1,}\\ \text{column 6)}}}{(-0.3636)}$$

$$p_{12} = 133.0909 + 0.0661$$
$$p_{12} = 133.1570$$

The computations of the six product moments are given in table 3.

The squared standard deviation, σ_1^2, was determined by the following formula:

$$\sigma_1^2 = AX_1^2 - AX_1AX_1 = AX_1^2 - (AX_1)^2$$

---

[9] A method of calculating sums of squares and sums of products with tabulating equipment is given in Appendix B, page 425.

The calculation was as follows:

$$\begin{aligned}\sigma_1^2 &= \underset{\substack{\text{(table 2,}\\ \text{column 1)}}}{198.3636} - \underset{\substack{\text{(table 1,}\\ \text{column 5)}}}{(0.1818)^2}\\ &= 198.3636 - 0.0331\\ &= 198.3305\end{aligned}$$

The computation of the four squared standard deviations is given in table 3.

Solution of the Simultaneous Equations

The values of the various combinations of the product sums and squared standard deviations from table 3 were substituted in the three normal equations

$$\begin{array}{ll}\text{(I)} & \sigma_2^2 b_{12.34} + p_{23} b_{13.24} + p_{24} b_{14.23} = p_{12}\\ \text{(II)} & p_{23} b_{12.34} + \sigma_3^2 b_{13.24} + p_{34} b_{14.23} = p_{13}\\ \text{(III)} & p_{24} b_{12.34} + p_{34} b_{13.24} + \sigma_4^2 b_{14.23} = p_{14}\end{array}$$

as follows:

$$\begin{array}{ll}\text{(I)} & 167.0496 b_{12.34} - 9.1570 b_{13.24} - 209.4300 b_{14.23} = 133.1570\\ \text{(II)} & -9.1570 b_{12.34} + 72.3078 b_{13.24} + 121.6713 b_{14.23} = -56.1033\\ \text{(III)} & -209.4300 b_{12.34} + 121.6713 b_{13.24} + 588.3984 b_{14.23} = -221.8304\end{array}$$

There are many ways of solving simultaneous equations. One of these methods is given in table 4.

The *first step* was the recording on lines 1, 2, and 3 the equations I, II, and III given above.

The *second step*: These three equations were divided by their respective coefficients of $b_{12.34}$. Line 4 was equation I divided through by the coefficient of $b_{12.34}$, which was 167.0496 (table 4). Line 5 was equation II divided by the coefficient of $b_{12.34}$, -9.1570. Line 6 was equation III divided by the coefficient of $b_{12.34}$, -209.4300.

The *third step* involved the determination of the successive differences between the three equations. Line 7 was line 4 minus line 5; and line 8 was line 5 minus line 6 (table 4). The work done to this point eliminated $b_{12.34}$, one of the three unknowns.

The *fourth step*:[10] The two equations in lines 7 and 8 were divided by their respective coefficients of $b_{13.24}$. Line 9 was line 7 divided by the coefficient of $b_{13.24}$, $+7.841635$; and line 10 was line 8 divided by the coefficient of $b_{13.24}$, -7.315487 (table 4).

The *fifth step*[11] involved the determination of the differences between

---

[10] Similar to the second step.

[11] Similar to the third step.

TABLE 4.—SOLUTION OF NORMAL EQUATIONS TO DETERMINE VALUES OF REGRESSION COEFFICIENTS

PRICE AND PRODUCTION OF WHEAT, TABLE 3

Line		Mechanical procedure	
1	Equation I	$167.0496b_{12.34} - 9.1570b_{13.24} - 209.4300b_{14.23} =$	133.1570
2	Equation II	$-9.1570b_{12.34} + 72.3078b_{13.24} + 121.6713b_{14.23} = -$	56.1033
3	Equation III	$-209.4300b_{12.34} + 121.6713b_{13.24} + 588.3984b_{14.23} = -$	221.8304
4	I÷167.0496	$b_{12.34} - 0.054816b_{13.24} - 1.253700b_{14.23} =$	0.797111
5	II÷−9.1570	$b_{12.34} - 7.896451b_{13.24} - 13.287245b_{14.23} =$	6.126821
6	III ÷ −209.4300	$b_{12.34} - 0.580964b_{13.24} - 2.809523b_{14.23} =$	1.059210
7	Line 4 −line 5	$7.841635b_{13.24} + 12.033545b_{14.23} = -$	5.329710
8	Line 5 −line 6	$-7.315487b_{13.24} - 10.477722b_{14.23} =$	5.067611
9	Line 7 ÷7.841635	$b_{13.24} + 1.534571b_{14.23} = -$	0.679668
10	Line 8 ÷ −7.315487	$b_{13.24} + 1.432266b_{14.23} = -$	0.692724
11	Line 9 −line 10	$0.102305b_{14.23} =$	0.013056
12	Line 11÷0.102305	$b_{14.23} =$	0.127618
13	Line 10 with value of $b_{14.23}$ substituted	$b_{13.24} + (1.432266)(0.127618) = -$	0.692724
14	Simplification	$b_{13.24} + 0.182783 = -$	0.692724
15	Value of $b_{13.24}$	$b_{13.24} = -$	0.875507
16	Line 6 with values of $b_{13.24}$ and $b_{14.23}$ substituted	$b_{12.34} - (0.580964)(-0.875507) - (2.809523)(0.127618) =$	1.059210
17	Simplification	$b_{12.34} + 0.508638 - 0.358546 =$	1.059210
18	Value of $b_{12.34}$	$b_{12.34} =$	0.909118

Check:

Equation I $167.0496b_{12.34} - 9.1570b_{13.24} - 209.4300b_{14.23} = 133.1570$

Substitute values of

$b_{12.34}, b_{13.24}, b_{14.23}$ $(167.0496)(0.909118) - (9.1570)(-0.875507) - (209.4300)(0.127618) = 133.1570$

$151.8678 + 8.0170 - 26.7270 = 133.1570$

$133.1578 = 133.1570$

$$R^2_{1.234} = \frac{b_{12.34}p_{12} + b_{13.24}p_{13} + b_{14.23}p_{14}}{\sigma^2_1}$$

$$= \frac{(0.909118)(133.1570) + (-0.875507)(-56.1033) + (0.127618)(-221.8304)}{198.3305}$$

$$= \frac{121.0554 + 49.1188 - 28.3096}{198.3305} = \frac{141.8646}{198.3305}$$

$$= 0.7153$$

$$R_{1.234} = 0.846$$

the two equations (lines 9 and 10). Line 11 was line 9 minus line 10. The work to this point eliminated both $b_{12.34}$ and $b_{13.24}$ and left only one unknown, $b_{14.23}$, in the equation (line 11). From this equation, which reads $+0.102305b_{14.23} = +0.013056$, it was possible to determine the value of $b_{14 \cdot 23}$ as follows:

$$b_{14.23} = \frac{+0.013056}{+0.102305} = +0.127618$$

This was given on line 12 (table 4).

The *sixth step* involved the substitution of the value of $b_{14.23}$, +0.127618, in the equation given in line 10, as follows:

$$b_{13.24} + 1.432266b_{14.23} = -0.692724$$

the computation of which appeared in lines 13, 14, and 15:

$$b_{13.24} + (1.432266)(+0.127618) = -0.692724$$
$$b_{13.24} + 0.182783 = -0.692724$$
$$b_{13.24} = -0.875507$$

The *seventh step* consisted of the substitution of the known values of $b_{14.23}$ and $b_{13.24}$ in the equation on line 6 (table 4).

$$b_{12.34} - 0.580964b_{13.24} - 2.809523b_{14.23} = +1.059210$$
$$b_{12.34} - (0.580964)(-0.875507) - (2.809523)(0.127618) = 1.059210$$
$$b_{12.34} + 0.508638 - 0.358546 = 1.059210$$
$$b_{12.34} = +0.909118$$

The values of the three unknowns have been determined by the solution of the normal equations.

The accuracy of the computations may be checked by substituting the computed values of $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$ in equation I (line 1, table 4).

$$167.0496b_{12.34} - 9.1570b_{13.24} - 209.4300b_{14.23} = 133.1570$$
$$(167.0496)(+0.909118) - (-9.1570)(-0.875507) - (209.4300)(+0.127618) = 133.1570$$
$$151.8678 + 8.0170 - 26.7270 = 133.1570$$
$$159.8848 - 26.7270 = 133.1570$$
$$133.1578 = 133.1570$$

Since the substitution of the values of the regression coefficients approximately satisfied the equation, it was reasonable to assume that no arithmetic errors had been made. It was possible, but not probable, that two or more errors which were compensating might have been made.

CALCULATION OF R

After the determination of the sums, product moments, standard deviations, and the regression coefficients, the student is able to determine the multiple correlation coefficient[12] from the following equation, which is carried forward from page 168:

[12] The calculation of R from any number of variables is given by the general formula,

$$R^2_{1.23\cdots m} = \frac{b_{12.34\cdots m}p_{12} + b_{13.24\cdots m}p_{13} + \cdots + b_{1m.23\cdots(m-1)}p_{1m}}{\sigma_1^2},$$

where m is the number of variables. The partial regression coefficients, $b_{12.34\ldots m}$ and the like, are obtained from $(m-1)$ normal equations of the type

$$\begin{array}{ccccccccc}
\sigma_2^2 b_{12.34\ldots m} & + & p_{23}b_{13.24\ldots m} & + & \cdots & + & p_{2m}b_{1m.23\ldots(m-1)} & = & p_{12} \\
p_{23}b_{12.34\ldots m} & + & \sigma_3^2 b_{13.24\ldots m} & + & \cdots & + & p_{3m}b_{1m.23\ldots(m-1)} & = & p_{13} \\
\cdot & & \cdot & & & & \cdot & & \cdot \\
\cdot & & \cdot & & & & \cdot & & \cdot \\
\cdot & & \cdot & & & & \cdot & & \cdot \\
p_{2m}b_{12.34\ldots m} & + & p_{3m}b_{13.24\ldots m} & + & \cdots & + & \sigma_m^2 b_{1m.23\ldots(m-1)} & = & p_{1m}
\end{array}$$

Coefficients of multiple correlation may also be obtained from the analysis of partial correlation (page 195).

$$R^2_{1.234} = \frac{b_{12.34}p_{12} + b_{13.24}p_{13} + b_{14.23}p_{14}}{\sigma_1^2}$$

$$R^2_{1.234} = 0.7153 \qquad R_{1.234} = 0.846$$

INTERPRETATION OF *R*

This multiple correlation coefficient, $R_{1.234} = 0.846$, indicated that there was a high degree of association between the Minneapolis price of wheat, X_1, and three factors: the Liverpool price, X_2; United States production, X_3; and world production, X_4.

The square of the coefficient, $R^2_{1.234} = 0.715$, indicated the proportion of the squared variability in the Minneapolis price of wheat explained by these three factors. The coefficient of determination was 0.715, or 71.5 per cent.

The unaccounted-for variability was expressed by the coefficient of non-determination, 0.285 $(1 - 0.715 = 0.285)$. The coefficient of non-determination was the proportion of the squared variability in the Minneapolis prices *not* explained by the three other factors.[13]

Simple correlation coefficients range from +1.0 to 0 to −1.0. The coefficients of multiple correlation are always positive in sign, and range from +1.0 to 0.

The multiple coefficient measured the combined effect of the three independent variables on the Minneapolis price, but gave no indication of the relative importance of the Liverpool price, or of the United States or world production. This problem will be treated under the subject of partial correlation.[14]

REGRESSION EQUATIONS

In the process of determining $R_{1.234}$, the values of the regression coefficients $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$ are obtained. The regression coefficient, $b_{12.34} = +0.9091$, indicated that, as the Liverpool price of wheat changed 1 cent per bushel, the Minneapolis price changed 0.9091 cent, in the same direction.[15] The subscripts to the letter b have an important meaning. The first two, "12," indicate that the coefficient describes the amount of change in X_1 with a unit change in X_2. The last two subscripts, "34," which are separated from the first two by a decimal point, indicate that the effects of X_3 and X_4 are eliminated in the determination of the relationship between X_1 and X_2. The sign of the regression coefficient $b_{12.34}$ is positive, indicating that the Minneapolis price rose and fell with the Liverpool price.

---

[13] This unaccounted-for variability might have been due to errors in the data on production, to errors in judgment as to what the market prices should have been, to other factors, or to inadequacy of the method.

[14] Page 185.

[15] The gross regression coefficient, b_{12} or b_{YX}, would also indicate the change in X_1 with a unit change in X_2, but the effects of X_3 and X_4 would not be considered.

The second regression coefficient, $b_{13.24} = -0.8755$, indicates that, with a change of 10 million bushels in the size of the United States wheat crop, the Minneapolis price changed 0.8755 cent in the opposite direction. In this relationship, the effects of the Liverpool price and world production are eliminated.

The third regression coefficient, $b_{14.23} = +0.1276$, indicates that, when the effects of the United States crop and Liverpool price are eliminated, a change of 10 million bushels in the world wheat crop is accompanied by a change of 0.1276 cent in the Minneapolis price. The two changes are in the same direction; normally one would expect that the changes would be in opposite directions.[16]

From these three regression coefficients, the regression equation for the estimated Minneapolis price in terms of these variables can be determined. This is a rather simple operation involving the three coefficients and the arithmetic averages of the four variables:

$$(X_1 - AX_1) = b_{12.34}(X_2 - AX_2) + b_{13.24}(X_3 - AX_3) + b_{14.23}(X_4 - AX_4)$$
$$(X_1 - 0.1818) = (0.909118)(X_2 + 0.3636) + (-0.875507)(X_3 - 0.3182) + (0.127618)(X_4 - 8.3182)$$
$$X_1 - 0.1818 = 0.9091X_2 + 0.3306 - 0.8755X_3 + 0.2786 + 0.1276X_4 - 1.0616$$
$$X_1 = 0.9091X_2 - 0.8755X_3 + 0.1276X_4 - 0.2706$$

This equation is the algebraic description of the average solid that best describes the multiple relationship. Solving for the value of X_1 in this equation for different values of X_2, X_3, and X_4 will give an estimate of the Minneapolis price based on the other factors. For instance, during the crop year 1907–1908, the Liverpool price, X_2, rose 16 cents; United States production, X_3, decreased 11 (tens of million bushels); and world production, X_4, declined 30 (tens of million bushels). When these values are substituted in the equation,

---

[16] This may be due to the elimination of two very important variables that influence the Minneapolis price. Since world production was probably a very important, if not dominating, factor influencing the Liverpool price, the Liverpool price was a reflection of world production. When the effect of the Liverpool price on the Minneapolis price was eliminated, the effect of world production was also partially or wholly removed. It might have been that changes in world wheat production affected the Minneapolis price through their effect on the Liverpool price.

There was some relationship between United States and world production, $r_{34} = +0.59$; and between United States production and the Minneapolis price, $r_{13} = -0.47$. By the elimination of the effect of the United States crop on the Minneapolis price, part of the effect of the world crop was also eliminated. The elimination of X_3 had some effect on $b_{12.34}$, but the elimination of X_4 was the more important.

$$X_1 = 0.9091X_2 - 0.8755X_3 + 0.1276X_4 - 0.2706$$
$$X_1 = (0.9091)(16) - (0.8755)(-11) + (0.1276)(-30) - 0.2706$$
$$X_1 = 14.5456 + 9.6305 - 3.8280 - 0.2706$$
$$X_1 = 20.0775$$

On the basis of these average relationships, the Minneapolis price should have risen 20 cents. It did rise 23 cents. Similar estimated values may be calculated for other years. An examination of a graphic comparison of the estimated and actual changes in prices indicates that a rather close relationship existed (figure 1). However, upon more detailed examination, it may be found that, in 1905, the estimated price declined 10 cents, while the actual price declined 27 cents. Likewise, the estimated price differed from the actual by 14 cents in 1912; 10 cents in 1895; and 8 to 9 cents in several other years. A comparison of estimated and actual prices over a period of time is one of the simplest and most commonly used graphic methods of showing the degree of correlation. In general, the reader who examines these graphs obtains the impression of a higher degree of correlation than is actually present. This is a criticism rather of graphic methods of showing relationships than of the multiple correlation analysis.

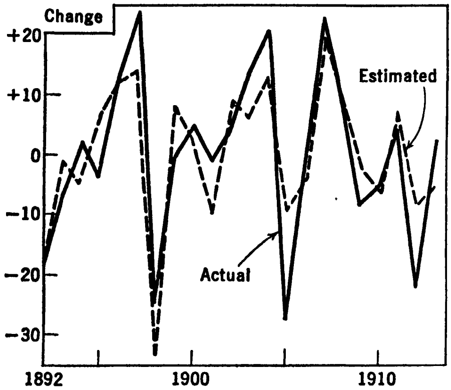


FIGURE 1.—ACTUAL AND ESTIMATED CHANGES IN THE MINNEAPOLIS PRICE OF WHEAT, 1892–1913

BASED ON MULTIPLE CORRELATION ANALYSIS
(Tables 1–4)

Most of the changes were in the same direction. The greatest discrepancies between actual and estimated changes were in the amount rather than the direction of change. These discrepancies are greater than appears from the usual visual inspection given this type of chart.

MULTIPLE CORRELATION FROM GRAPHIC ANALYSIS

Multiple correlation coefficients and regression lines may be obtained by a short-cut graphic method. This method involves less work than the least-squares method but is also less accurate. The short-cut method is described on pages 230 to 241.

ADVANTAGES OF MULTIPLE CORRELATION ANALYSIS

The greatest advantage of multiple correlation over other methods of studying association between variables lies in its adaptability to problems where the amount of data is relatively small.

If the records were available since 1700, there would be about 240 years of varying combinations of Minneapolis and Liverpool prices and United States and world production. These data might be sorted and subsorted in various ways to determine the Minneapolis price with various combinations of the Liverpool price and United States and world production. If the years were sorted on the basis of the Liverpool price, X_2, into two groups, one in which the prices rose, and the other in which the prices fell, there would be about 120 years in each group. If each of these two groups was again divided on the basis of whether United States production increased or decreased, there would be four subgroups of about 60 years each.[17] If these four subgroups were divided on the basis of whether world production increased or decreased, there would be eight sub-subgroups[17] with about 30 years each. For each group of 30 years, the average change in the Minneapolis price could be calculated and compared with the average changes for other groups. This type of analysis would be much simpler and less time-consuming than multiple correlation analysis.of the same data and probably would be about as satisfactory, except for the complexity of the results.

In the 22-year period in table 1, there would not be more than two or three years, on the average, in each of the eight subgroups.[18] While the averages of 30 items might be reasonably reliable, averages based on two or three items would not. Multiple correlation is also an averaging process. However, it gives reasonably accurate results with limited amounts of data, whereas the averaging of groups and subgroups does not.

Another advantage of multiple correlation analysis is the expression of the type and degree of relationship in a few concise coefficients. To state that $R^2_{1.234} = 0.715$ means that the three factors represented, X_2, X_3, and X_4, explain 71.5 per cent of the squared variability in X_1. The regression coefficients are likewise well defined and have a definite meaning.

The reliability of the correlation and regression coefficients is more easily tested than the reliability of other methods of studying association.

Compared with other types of multiple correlation, linear multiple analysis is relatively simple.

---

[17] Provided that there was no interrelationship between the various factors. Since there was interrelationship between all pairs of factors in the example given, equal distribution of the years among the subgroups could not be expected.

[18] When some interrelationship existed among three factors, some of these groups would probably contain no years at all. In fact, in the example used, it never happened that the Liverpool price declined, the United States crop increased, and the world crop decreased all in the same year. In six of the years, the Liverpool price declined and the United States and world crops increased all in the same year.

DISADVANTAGES OF MULTIPLE CORRELATION ANALYSIS

Multiple correlation analysis is based on the assumption that the relationships between the variables are linear. In other words, the rate of change in one variable in terms of another is assumed to be constant for all values. In the field of agriculture, most relationships are not linear but follow some other pattern. This somewhat limits the use of multiple correlation analysis.[19] The linear regression coefficients are not accurately descriptive of curvilinear data.

A second important disadvantage or limitation is the assumption that the effects of independent variables on the dependent variables are separate, distinct, and additive. When the effects of variables are additive, a given change in one has the same effect on the dependent variable regardless of the sizes of the other two independent variables. For example, in the equation

$$X_1 = 0.9091X_2 - 0.8755X_3 + 0.1276X_4 - 0.2706$$

the Minneapolis price of wheat, X_1, increased 0.9091 cent with every 1-cent increase in the Liverpool price, X_2, regardless of the sizes of production in the United States and in the world. However, the effect of the Liverpool price on the Minneapolis price may be different when the United States production is high from that when it is low. It often happens in agricultural data that the effect of the first factor upon the dependent variable is reversed with a change in the size of the second or third factor. When the effects of any variable change with different values of another variable, their two effects are not additive, but are said to be joint.

In the multiple regression equation, the various terms, which are products of regression coefficients and the independent variables, are added to each other. Often the equation which would give the best fit would contain terms in quite different combinations. Such a combination with four variables is given in the following equation: $X_1 = \frac{b_{12.34}X_2}{b_{13.24}X_3 - a} + b_{14.23}X_4$. The value of the first term, $\frac{b_{12.34}X_2}{b_{13.24}X_3 - a}$, depends on both X_2 and X_3. The effect of changes in the value of X_2 on X_1 depends on the value of X_3. When X_3 is at one level, an increase in X_2 may mean an *increase* in X_1; when X_3 is at another level, an increase in X_2 could mean a *decrease* in X_1. Therefore, the relation of

---

[19] When linear correlation methods are applied to curvilinear data, the degree of relationship is really greater than that indicated by the coefficient of correlation. However, a small departure from linearity does not seriously affect the results.

X_2 and X_3 to X_1 is joint. Such joint relationships of independent variables to the dependent may take the form of products, quotients, powers, roots, and other complicated functions.

Multiple correlation analysis assumes the simplest of the possible interrelationships among the independent variables, namely, the additive relationship. Often this assumption does not agree with fact.

The method has several other disadvantages. Although it is less laborious than most curvilinear correlation analyses, linear multiple correlation involves a great deal of work relative to the results frequently obtained. When the results are obtained, only a few students well trained in the method are able to interpret them. The misuse of correlation results has probably cast more doubt on the method than is justified. However, this lack of understanding and resulting misuse are due to the complexity of the method and are thereby disadvantages chargeable to it.

USES OF MULTIPLE CORRELATION COEFFICIENTS

Multiple correlation is used in many fields of experimental endeavor. Hitchcock found that the labor cost of producing maple syrup and sugar in Vermont was related to the size of the orchard, the sugar content of the sap, and the yield per bucket (table 5). The multiple correlation, $R = 0.50$, indicated that these factors accounted for about 25 per cent of the variability.

Kincer and Mattice studied variations in the yield of spring wheat in North Dakota from 1900 to 1924. The independent variables considered were the amount of sunshine in July and the total rainfall in April, May, and June. The relationship was rather high, as indicated by the coefficient $R = 0.80$.

Vial studied the relation of retail price to the content of fertilizers. For the year 1902, he found that the nitrogen, phosphoric acid, and potash contents were closely related to the price, $R = 0.88$. Substantially the same results were obtained for each of the 39 years studied. The coefficients ranged from 0.75 to 0.96.

Cox studied the relation of the local and Corn Belt production to the Minnesota farm price of corn. The coefficient, $R = 0.83$, indicated that these production factors explained almost 70 per cent of the variations in the Minnesota price.

The multiple correlation coefficient has been used quite extensively to measure the degree of association between variables. This coefficient has been used more frequently than its square, R^2, which indicates the percentage of determination. However, R^2 has more meaning and is preferable.

TABLE 5.—MULTIPLE CORRELATION STUDIES

Dependent variable	Independent variables		Multiple correlation coefficient	Dependent variable	Independent variables		Multiple correlation coefficient
	Number	Description			Number	Description	
Maple syrup[*] Man labor per gallon	3	Size of orchard; sugar content; yield per bucket	0.50	*Fertilizer*[¶] Retail price	3	Nitrogen; phosphoric acid; potash	0.88
Spring wheat[†] Yield, No. Dakota	2	July sunshine; April-June rainfall	0.80	*Cotton*[**] Acreage	4	Time; price one and two years preceding; acreage preceding year	0.95
Milk[‡] Deliveries	2	Milk-feed price ratio lagged 1-½ and 12-½ months	0.65	*Land*[††] Value	4	Class; productivity; buildings; market	0.81
Potatoes[§] Idaho price	4	Production in Far West, in Central States, in Far East; price level	0.97	*Rice*[‡‡] Price	2	United States supply; India production	0.97
Eggs[‖] Retail price	4	Weight; quality; cleanliness; type of store	0.48	*Corn*[§§] Minnesota price	2	Production in Minnesota; and in six Corn Belt states	0.83

---

[*] Hitchcock, J. A., Economics of the Farm Manufacture of Maple Syrup and Sugar, Vermont Agricultural Experiment Station, Bulletin 285, p. 64, July 1928.

[†] Kincer, J. B., and Mattice, W. A., Statistical Correlations of Weather Influence on Crop Yields, Monthly Weather Review, Vol. 56, p. 56, February 1928.

[‡] Gans, A. R., Elasticity of Supply of Milk from Vermont Plants, Vermont Agricultural Experiment Station, Bulletin 269, p. 27, April 1927.

[§] Heflebower, R. B., Factors Relating to the Price of Idaho Potatoes, University of Idaho Agricultural Experiment Station, Bulletin 166, p. 15, June 1929.

[‖] Benner, C. L., and Gabriel, H. S., Marketing of Delaware Eggs, University of Delaware Agricultural Experiment Station, Bulletin 150, p. 23, July 1927.

[¶] Vial, E. E., Retail Prices of Fertilizer Materials and Mixed Fertilizers, Cornell University Agricultural Experiment Station, Bulletin 545, p. 120, November 1932.

[**] Smith, B. B., Forecasting the Volume and Value of the Cotton Crop, Journal of the American Statistical Association, Vol. XXII, New Series, No. 160, p. 449, December 1927.

[††] Haas, G. C., Sale Prices as a Basis for Farm Land Appraisal, University of Minnesota Agricultural Experiment Station, Technical Bulletin 9, p. 22, November 1922.

[‡‡] Campbell, C. E., Factors Affecting the Price of Rice, United States Department of Agriculture, Technical Bulletin 297, p. 21, April 1932.

[§§] Cox, R. W., Factors Influencing Corn Prices, University of Minnesota Agricultural Experiment Station, Technical Bulletin 81, p. 18, September 1931.

USES OF MULTIPLE REGRESSION EQUATIONS

The regression coefficients obtained by multiple correlation analysis are in many instances more informative than the correlation coefficients. They give the average rates of change in one variable with changes in a second variable, the influence of other variables being eliminated.

Bennett[20] found that the price of No. 2 oats at Chicago was closely associated with the supply of both corn and oats. The equation for the November–April price of No. 2 oats for the 42-year period 1897 to 1938 was: $X_1 = 237.4 - 0.903X_2 - 0.471X_3$ where X_2 was the August 1 supply of oats; and X_3, the November 1 supply of corn. This equation indicated that, for each increase of 1 per cent in the supply of oats, the price fell 0.90 per cent. With an increase of 1 per cent in the supply of corn, the price declined 0.47. These relations can be shown in tabular form. If one were interested in the effect of the supply of oats on the price of oats with the supply of corn eliminated, this relationship could be obtained from the net relationship given above. The relation of the supply of oats to its price with the effect of the supply of corn eliminated may be obtained from the above equation by holding constant the supply of corn, X_3, at its average,[21] 100. From the resulting equation, $X_1 = 190.3 - 0.903X_2$, the price of oats may be determined for different supplies of oats. For example, when the supply of oats was 20 per cent below normal, the price was 18 per cent above normal[22] (table 6).

TABLE 6.—EFFECT OF SUPPLY OF OATS ON PRICE OF OATS, WITH THE EFFECT OF THE SUPPLY OF CORN ELIMINATED

SUPPLIES AND PURCHASING POWER OF PRICE IN PERCENTAGE OF NORMAL*

Supply oats, X_2	Price of oats, X_1
80	118
90	109
100	100
110	91
120	82

* Based on the equation $X_1 = 190.3 - 0.903X_2$.

If one were interested in the effect of the supply of corn on the price of oats, with the effect of the supply of oats eliminated, the same procedure would be followed. When corn was 20 per cent below normal, the price of oats was 9 per cent above normal (table 7).

A shortage of 20 per cent in the oat supply, with the effect of the corn supply eliminated, raised oat prices 18 per cent. A similar shortage in the corn supply, with the effect of the oat supply eliminated, raised prices 9 per cent. Obviously, the oat supply had a greater effect on the price of oats than did the corn supply.

---

[20] Bennett, K. R., The Price of Feed, unpublished manuscript, Cornell University, 1940.

[21] $X_1 = 237.4 - 0.903X_2 - 0.471(100)$.
$X_1 = 237.4 - 47.1 - 0.903X_2$.
$X_1 = 190.3 - 0.903X_2$.

[22] This assumes that the corn supply was average.

TABLE 7.—EFFECT OF SUPPLY OF CORN ON PRICE OF OATS, WITH THE EFFECT OF THE SUPPLY OF OATS ELIMINATED

SUPPLIES AND PURCHASING POWER OF PRICE IN PERCENTAGE OF NORMAL*

Supply corn, X_3	Price of oats, X_1
80	109
90	105
100	100
110	95
120	91

* Based on the equation $X_1 = 147.1 - 0.471X_3$.

The combined effect of various supplies of oats and corn can also be shown in tabular form. The average prices for these combinations can be easily determined from the original equation of relationship (table 8). With the supply of both oats and corn 20 per cent below normal, the price was 27 per cent above normal.[23] When the oat supply was 10 per cent below normal and the corn supply 20 per cent above normal, the price of oats was normal. If one so desired, the supply of oats and corn could be converted to bushels; and the prices, to cents per bushel. In any event, this type of presentation is more effective than the regression equation or its coefficients. However, the tabular presentation does not overcome the inherent limitations of multiple correlation analysis.

TABLE 8.—EFFECT OF SUPPLY OF OATS AND OF CORN ON PRICE OF OATS

SUPPLIES AND PURCHASING POWER OF PRICE IN PER CENT OF NORMAL

Supply oats, X_2	Supply corn, X_3				
	80	90	100	110	120
	Price oats, X_1*	*Price oats,* X_1*	*Price oats,* X_1*	*Price oats,* X_1*	*Price oats,* X_1*
80	127	123	118	113	109
90	118	114	109	104	100
100	109	105	100	95	91
110	100	96	91	86	82
120	91	87	82	77	73

* Based on the equation $X_1 = 237.4 - 0.903X_2 - 0.471X_3$.

[23] $X_1 = 237.4 - 0.903(80) - 0.471(80)$.
$X_1 = 237.4 - 72.2 - 37.7$.
$X_1 = 127.5$.

CHAPTER 11

PARTIAL CORRELATION

Multiple correlation measures the degree of relation between one variable, such as price, production, income, and the like, and a combination of two or more other, related factors. However, it tells one nothing about the relative importance of each factor. In statistical analysis, the worker is usually not satisfied in measuring only the combined effect of all factors. If enough factors were included in the multiple correlation and no errors of any kind were made, the multiple coefficient would always approach 1.0. The usual coefficient of less than 1.0 simply measures the degree to which the student has found the causes of variation in the dependent variable; and, for this purpose, it is a useful tool. But after the multiple correlation is known, the student focuses his attention on the relative importance of each factor considered. This problem is of greatest importance if for no other reason than to clear up the ambiguity of the multiple coefficient. A high coefficient of multiple correlation connotes to the average student a high degree of association between the dependent variable and each of the independent variables. In many cases, however, the high coefficient may be due to high association between only one or two of the factors. The other factors may be of practically no importance.

PARTIAL CORRELATION FROM MULTIPLE CORRELATIONS

SECOND-ORDER COEFFICIENTS

The partial correlation coefficient is a measure of the effect of one factor on the dependent variable when the effects of all the other factors considered are eliminated. One of the best definitions of the partial correlation coefficient was given by Ezekiel[1]: "The coefficient of partial correlation may be defined as a measure of the extent to which that part of the variation in the dependent variable which was *not* explained by the other independent factors can be explained by the addition of the new factor."

With every additional independent variable in a correlation problem, the multiple coefficient increases, or remains the same as before. If the increase is large, the effect of the additional variable is important. If there is no increase, its effect is negligible. The partial correlation coefficient compares this increase in the multiple coefficient to the proportion of the variability in the dependent variable not explained by the factors first considered. The increase in the per cent determination due to the inclusion of a third independent variable might be given by the expression $R^2_{1.234} - R^2_{1.23}$. The proportion of variability in X_1 not explained by X_2 and X_3 was $1 - R^2_{1.23}$. The ratio of the increase in accounted-for variability due to the inclusion of X_4 to the proportion not explained by X_2 and X_3 is the partial correlation coefficient squared. The value of the partial coefficient[2] is thus given:

---

[1] Ezekiel, M., Methods of Correlation Analysis, p. 179, 1930.

$$r_{14.23} = \sqrt{\frac{R^2_{1.234} - R^2_{1.23}}{1 - R^2_{1.23}}}$$

This can be interpreted as the effect of X_4 on X_1 after the effects of X_2 and X_3 have been eliminated.

Likewise, the partial coefficient, $r_{12.34}$, would be given by:

$$r_{12.34} = \sqrt{\frac{R^2_{1.234} - R^2_{1.34}}{1 - R^2_{1.34}}} \quad \text{and} \quad r^2_{12.34} = \frac{R^2_{1.234} - R^2_{1.34}}{1 - R^2_{1.34}}$$

It was desired to measure the effect of the Liverpool price, X_2, on the Minneapolis price of wheat, X_1, *with the effects of United States and world production, X_3 and X_4, eliminated.* This is measured by the partial correlation coefficient, $r_{12.34}$:

$$r^2_{12.34} = \frac{0.7153 - 0.4329}{1 - 0.4329} = \frac{0.2824}{0.5671} = 0.4980$$

$$r_{12.34} = +0.706$$

All the multiple correlation coefficients and the regression coefficients on which the partial coefficients are based are summarized in table 1.

The effect of the Liverpool price and the United States and world production on the Minneapolis price of wheat was measured by $R_{1.234} = 0.846$ and $R^2_{1.234} = 0.7153$ (table 1). These three factors explained 71.53 per cent of the squared variability in the Minneapolis price. However, consideration of only two of the three independent variables, United States and world production, accounted for only 43.29 per cent of the variability[3] in the Minneapolis price ($R^2_{1.34} = 0.4329$). The inclu-

---

[2] This formula usually appears in the following transformation:

$$r_{14.23} = \sqrt{1 - \frac{(1 - R^2_{1.234})}{(1 - R^2_{1.23})}}$$

[3] Computed by formula: $R^2_{1.34} = \frac{b_{13.4}p_{13} + b_{14.3}p_{14}}{\sigma_1^2}$. The general formula for multiple coefficients with any number of variables is given in footnote 12, page 175.

sion of the additional factor, Liverpool price, X_2, raised the per cent determination from 43.29 to 71.53. The increase, 28.24 per cent, was expressed as a proportion of the variability in X_1 not explained by X_3 and X_4, 56.71 per cent (1 − 0.4329 = 0.5671). This proportion, 0.4980, is the squared partial correlation coefficient $\left(r^2_{12.34} = \frac{0.2824}{0.5671} = 0.4980\right)$. The coefficient itself, $r_{12.34} = +0.706$, is the square root of this quantity.

TABLE 1.—SECOND-ORDER PARTIAL CORRELATION COEFFICIENTS AND THE MULTIPLE AND REGRESSION COEFFICIENTS FROM WHICH THEY ARE DERIVED

PRODUCTION AND PRICE OF WHEAT*

Multiple coefficients		Regression coefficients	Partial correlation coefficients	Corresponding gross coefficients
Four variables	Three variables			
$R_{1.234} = 0.846$	$R_{1.34} = 0.658$	$b_{12.34} = +0.9091$	$r_{12.34} = +0.706$	$r_{12} = +0.732$
$R^2_{1.234} = 0.7153$	$R^2_{1.34} = 0.4329$		$r^2_{12.34} = 0.498$	$r^2_{12} = 0.535$
$R_{1.234} = 0.846$	$R_{1.24} = 0.763$	$b_{13.24} = -0.8755$	$r_{13.24} = -0.565$	$r_{13} = -0.469$
$R^2_{1.234} = 0.7153$	$R^2_{1.24} = 0.5818$		$r^2_{13.24} = 0.3192$	$r^2_{13} = 0.220$
$R_{1.234} = 0.846$	$R_{1.23} = 0.838$	$b_{14.23} = +0.1276$	$r_{14.23} = +0.208$	$r_{14} = -0.649$
$R^2_{1.234} = 0.7153$	$R^2_{1.23} = 0.7024$		$r^2_{14.23} = 0.0433$	$r^2_{14} = 0.422$

* Table 1, page 170.

The partial coefficient, $r_{12.34}$, takes the sign of the regression coefficient, $b_{12.34}$. In the derivation of the multiple correlation coefficient, $R_{1.234}$, this regression was found to be positive,[4] $b_{12.34} = +0.9091$.

This partial coefficient, $r_{12.34} = +0.706$, measures the extent[5] to which the variation in the Minneapolis price, unexplained by United States and world production, was explained by the Liverpool price.

Partial coefficients, though logically derived from multiple coefficients, may be compared with the gross, or simple, correlation coefficients. The gross correlation between the Liverpool and Minneapolis prices was $r_{12} = +0.732$ (table 1). This can be interpreted as the relation between X_1 and X_2 without considering X_3 and X_4. The partial coefficient, $r_{12.34} = +0.706$, is interpreted as the relation between X_1 and X_2 *after the effects of* X_3 *and* X_4 *have been eliminated.* In our problem, the gross and partial coefficients were about the same. The elimination of the effects of X_3 and X_4 decreased the correlation coefficient only slightly. This indicated a persistent relation between the Minneapolis and Liverpool prices, regardless of world or United States production.

---

[4] Calculations in table 4, page 174.

[5] The coefficient squared, $r^2_{12.34} = 0.498$, measures the proportion of the variation in the Minneapolis price, unexplained by the United States and world production, which was explained by the Liverpool price.

The partial correlation between the Minneapolis price and the United States production, with the effects of world production and Liverpool prices eliminated, was $r_{13.24} = -0.565$. This coefficient was calculated from the squares of the multiple coefficients, $R^2_{1.234} = 0.7153$ and $R^2_{1.24} = 0.5818$, as follows:

$$r_{13.24} = \sqrt{\frac{R^2_{1.234} - R^2_{1.24}}{1 - R^2_{1.24}}} = \sqrt{\frac{0.7153 - 0.5818}{1 - 0.5818}} = \sqrt{\frac{0.1335}{0.4182}} = \sqrt{0.3192} = -0.565$$

Since $b_{13.24}$ was negative,[6] the sign of the partial correlation was also negative, $r_{13.24} = -0.565$. The gross correlation between United States production, X_3, and the Minneapolis price, X_1, was $r_{13} = -0.469$ (table 1). The elimination of the effects of world production, X_4, and the Liverpool price, X_2, actually raised the coefficient somewhat. This indicated that the apparent effects of the United States production, $r_{13} = -0.469$, really were the effects of the United States production and not merely a reflection of the other two factors, which were related to both United States production and the Minneapolis price. The square of the partial correlation coefficient, $r^2_{13.24} = 0.319$, indicated that the United States production accounted for 31.9 per cent of that part of the variability in the Minneapolis price not accounted for by world production and the Liverpool price.

Similarly, the partial correlation between world production and the Minneapolis price may be calculated from the squared multiple coefficients, $R^2_{1.234} = 0.7153$ and $R^2_{1.23} = 0.7024$. This partial coefficient,[7] $r_{14.23} = +0.208$, indicated that there was little correlation between[8] world production and the Minneapolis price, in addition to that already reflected through the effects of the Liverpool price and the United States production on the Minneapolis price. An examination of the squared partial coefficient, $r^2_{14.23} = 0.043$, reveals that world production explained practically none of the variability in the Minneapolis price not already explained by the other two factors.

The three partial correlation coefficients, $r_{12.34}$, $r_{13.24}$, and $r_{14.23}$, are known as second-order partials, there being in each case two independent variables whose effects were eliminated.[9]

---

[6] $b_{13.24} = -0.8755$ (table 1).

[7] $b_{14.23} = +0.1276$ (table 1).

[8] The coefficient $r_{14.23} = +0.208$ was undoubtedly not significant. The gross coefficient, $r_{14} = -0.649$, was much greater and had a different sign from its corresponding partial, $r_{14.23} = +0.208$.

TABLE 2.—FIRST-ORDER PARTIAL CORRELATION COEFFICIENTS AND THE MULTIPLE, GROSS, AND REGRESSION COEFFICIENTS FROM WHICH THEY ARE DERIVED

PRODUCTION AND PRICE OF WHEAT

Multiple and gross coefficients		Regression coefficients	Partial correlation coefficients	Corresponding gross coefficients
Three variables	Two variables			
$R_{1.23} = 0.838$	$r_{13} = -0.469$	$b_{12.3} = +0.760$	$r_{12.3} = +0.787$	$r_{12} = +0.732$
$R^2_{1.23} = 0.7024$	$r^2_{13} = 0.2195$		$r^2_{12.3} = 0.6187$	$r^2_{12} = 0.535$
$R_{1.24} = 0.763$	$r_{14} = -0.649$	$b_{12.4} = +0.586$	$r_{12.4} = +0.526$	$r_{12} = +0.732$
$R^2_{1.24} = 0.5818$	$r^2_{14} = 0.4217$		$r^2_{12.4} = 0.2768$	$r^2_{12} = 0.535$
$R_{1.23} = 0.838$	$r_{12} = +0.732$	$b_{13.2} = -0.680$	$r_{13.2} = -0.600$	$r_{13} = -0.469$
$R^2_{1.23} = 0.7024$	$r^2_{12} = 0.5352$		$r^2_{13.2} = 0.3597$	$r^2_{13} = 0.220$
$R_{1.34} = 0.658$	$r_{14} = -0.649$	$b_{13.4} = -0.217$	$r_{13.4} = -0.139$	$r_{13} = -0.469$
$R^2_{1.34} = 0.4329$	$r^2_{14} = 0.4217$		$r^2_{13.4} = 0.0194$	$r^2_{13} = 0.220$
$R_{1.24} = 0.763$	$r_{12} = +0.732$	$b_{14.2} = -0.168$	$r_{14.2} = -0.317$	$r_{14} = -0.649$
$R^2_{1.24} = 0.5818$	$r^2_{12} = 0.5352$		$r^2_{14.2} = 0.1003$	$r^2_{14} = 0.422$
$R_{1.34} = 0.658$	$r_{13} = -0.469$	$b_{14.3} = -0.332$	$r_{14.3} = -0.523$	$r_{14} = -0.649$
$R^2_{1.34} = 0.4329$	$r^2_{13} = 0.2195$		$r^2_{14.3} = 0.2734$	$r^2_{14} = 0.422$

FIRST-ORDER COEFFICIENTS

First-order partial coefficients measure the relationship between two variables with the effect of a third eliminated. They can be calculated from the multiple correlation coefficients involving three variables and the gross coefficients. The correlation between the Liverpool and Minneapolis prices, with only the effect of United States production eliminated, was calculated from $R^2_{1.23} = 0.7024$ and $r^2_{13} = 0.2195$ (table 2).

$$r_{12.3} = \sqrt{\frac{R^2_{1.23} - r^2_{13}}{1 - r^2_{13}}} = \sqrt{\frac{0.7024 - 0.2195}{1 - 0.2195}} = \sqrt{\frac{0.4829}{0.7805}} = \sqrt{0.6187} = +0.787$$

This partial coefficient takes the sign of the regression coefficient[10] and reads $r_{12.3} = +0.787$. The prior elimination of the effect of the United

---

[9] General formula for the calculation of any order partials for the X_1X_2 relationship from multiple coefficients is

$$r_{12.34\cdots m} = \sqrt{\frac{R^2_{1.234\cdots m} - R^2_{1.34\cdots m}}{1 - R^2_{1.34\cdots m}}}$$

The formulas for $r_{13.24 \ldots m}$, $r_{14.23 \ldots m}$, etc., may be obtained by interchanging 3 and 2, 4 and 2, etc., in the above formula.

[10] $b_{12.3} = +0.760$ (table 2).

States production, X_3, upon the Minneapolis price, X_1, raised somewhat the correlation between the Minneapolis and Liverpool prices, $r_{12} = +0.732$ and $r_{12.3} = 0.787$. While the Liverpool price accounted for about 54 per cent of the total variability[11] in the Minneapolis price, it explained 62 per cent of this variability not accounted for by United States production.

COMPARISON OF PARTIAL AND GROSS COEFFICIENTS

The data necessary to calculate the second- and first-order partial correlation coefficients are given in tables 1 and 2, respectively. The last column contains the gross coefficients and their respective squares with which the partials or their squares (next to last column) may be compared.

A comparison of the partial with the gross correlation coefficients reveals that the elimination of other independent variables sometimes reduced the coefficient greatly and sometimes did not materially change it one way or the other. If the partial was about the same as the gross coefficient, a persistent and separate relationship to the Minneapolis price was indicated; that is, the relationship was not through one of the other independent factors.

The partial, $r_{12.34} = +0.706$, was about the same as the gross, $r_{12} = +0.732$. There was a persistent and separate relationship between the Minneapolis and Liverpool prices of wheat, X_1 and X_2, in addition to that expressed through the mutual relation of each to United States and world production, X_3 and X_4.

If the partial was much less than the gross coefficient, there was interrelationship between this independent variable and another which itself was related to the Minneapolis price. This principle was clearly demonstrated by a comparison of the partial and gross correlations between the Minneapolis price, X_1, and world production, X_4, $r_{14.23} = +0.208$ and $r_{14} = -0.649$. The striking change was due to interrelationship between world production, X_4, and Liverpool price, X_2 ($r_{24} = -0.668$), and between United States production, X_3, and world production, X_4 ($r_{34} = +0.590$). Before consideration of world production, the Liverpool price and the United States production had already explained 70.2 per cent of the variability in the Minneapolis price, and world production explained little of the variation that was left, $r^2_{14.23} = 0.043$.

---

[11] The gross correlation, $r_{12} = 0.732$, and the partial, $r_{12.3} = 0.787$, when squared were 0.535 and 0.619, respectively.

PARTIAL ANALYSIS FROM GROSS CORRELATIONS

Correlation analysis may proceed (*a*) from multiple and gross to partial coefficients; or (*b*) from gross to partial to multiple coefficients. The calculation of partial correlation proceeding from multiples to partials[12] has already been explained. The calculation of partial coefficients proceeding from gross coefficients, which has been historically important, follows.

FIRST-ORDER COEFFICIENTS

If one were interested in the partial correlation between the two variables X_1 and X_2, *with the effect of a third variable*, X_3, *eliminated*, the coefficient $r_{12.3}$ could be determined from the three gross correlations r_{12}, r_{13}, and r_{23} by the following formula:

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

The gross coefficient between the Minneapolis and Liverpool prices of wheat, X_1 and X_2, was $r_{12} = +0.7316$; between the Minneapolis price, X_1, and the United States production, X_3, $r_{13} = -0.4685$; and between the Liverpool price, X_2, and the United States production, X_3, $r_{23} = -0.0833$ (table 3). The first-order partial correlation between the Minneapolis and Liverpool prices, with United States production eliminated, was:[13]

$$\begin{aligned} r_{12.3} &= \frac{0.7316 - (-0.4685)(-0.0833)}{\sqrt{1 - (-0.4685)^2}\sqrt{1 - (-0.0833)^2}} \\ &= \frac{0.7316 - 0.0390}{(0.8835)(0.9965)} \\ &= \frac{0.6926}{0.8804} = +0.7867 \end{aligned}$$

The sign of the partial coefficient calculated by this method was determined by the net value of the terms in the numerator of the formula. The partial coefficient, $r_{12.3} = +0.787$, was a measure of the relation between Minneapolis and Liverpool prices of wheat, X_1 and X_2, with the effect of the United States production, X_3, eliminated (table 3). The partial coefficient developed from gross coefficients was identical, in value and in interpretation, with that developed from multiple coefficients (compare tables 2 and 3).

[12] Tables 1 and 2, pages 187 and 189.

[13] The values of the expression $\sqrt{1 - r^2}$ are somewhat difficult to calculate but can be easily read from Miner, J. R., Tables of $\sqrt{1 - r^2}$ and $1 - r^2$, 1922.

TABLE 3.—DETERMINATION OF PARTIAL CORRELATION COEFFICIENTS FROM LOWER-ORDER COEFFICIENTS

PRICE AND PRODUCTION OF WHEAT*

Correlation coefficient		Product term of numerator	Whole numerator	$\sqrt{1-r^2}$	Denominator	Partial	
Subscript	Coefficient					Subscript	Coefficient
Calculation of first-order partial coefficients†							
r_{12}	+ 0.7316	+ 0.0390	+ 0.6926		0.8804	$r_{12.3}$	+ 0.7867
r_{13}	− 0.4685			0.8835			
r_{23}	− 0.0833			0.9965			
r_{12}	+ 0.7316	+ 0.4338	+ 0.2978		0.5659	$r_{12.4}$	+ 0.5262
r_{14}	− 0.6494			0.7604			
r_{24}	− 0.6680			0.7442			
r_{13}	− 0.4685	− 0.0609	− 0.4076		0.6793	$r_{13.2}$	− 0.6000
r_{12}	+ 0.7316			0.6817			
r_{23}	− 0.0833			0.9965			
r_{14}	− 0.6494	− 0.2764	− 0.3730		0.7134	$r_{14.3}$	− 0.5228
r_{13}	− 0.4685			0.8835			
r_{34}	+ 0.5899			0.8075			
r_{14}	− 0.6494	− 0.4887	− 0.1607		0.5073	$r_{14.2}$	− 0.3168
r_{12}	+ 0.7316			0.6817			
r_{24}	− 0.6680			0.7442			
r_{13}	− 0.4685	− 0.3831	− 0.0854		0.6140	$r_{13.4}$	− 0.1391
r_{14}	− 0.6494			0.7604			
r_{34}	+ 0.5899			0.8075			
r_{23}	− 0.0833	− 0.3941	+ 0.3108		0.6009	$r_{23.4}$	+ 0.5172
r_{24}	− 0.6680			0.7442			
r_{34}	+ 0.5899			0.8075			
r_{34}	+ 0.5899	+ 0.0556	+ 0.5343		0.7416	$r_{34.2}$	+ 0.7205
r_{23}	− 0.0833			0.9965			
r_{24}	− 0.6680			0.7442			
Calculation of second-order partial coefficients†							
$r_{12.3}$	+ 0.7867	+ 0.4021	+ 0.3846		0.5448	$r_{12.34}$	+ 0.7059
$r_{14.3}$	− 0.5228			0.8525			
$r_{24.3}$	− 0.7691			0.6391			
$r_{13.2}$	− 0.6000	− 0.2283	− 0.3717		0.6578	$r_{13.24}$	− 0.5651
$r_{14.2}$	− 0.3168			0.9485			
$r_{34.2}$	+ 0.7205			0.6935			
$r_{14.2}$	− 0.3168	− 0.4323	+ 0.1155		0.5548	$r_{14.23}$	+ 0.2082
$r_{13.2}$	− 0.6000			0.8000			
$r_{34.2}$	+ 0.7205			0.6935			

---

* Calculated from table 1, page 170.

† There are 8 first-order and 3 second-order partial coefficients presented. There are 12 possible first-order and 6 possible second-order coefficients.

(*Footnote continued on page 193*)

The partial relationship between the Minneapolis price, X_1, and United States production, X_3, with world production, X_4, eliminated, was:

$$
\begin{aligned}
r_{13.4} &= \frac{r_{13} - (r_{14})(r_{34})}{\sqrt{1 - r_{14}^2}\sqrt{1 - r_{34}^2}} \\
&= \frac{-0.4685 - (-0.6494)(+0.5899)}{\sqrt{1 - (-0.6494)^2}\sqrt{1 - (0.5899)^2}} \\
&= \frac{-0.4685 + 0.3831}{(0.7604)(0.8075)} \\
&= \frac{-0.0854}{0.6140} = -0.1391
\end{aligned}
$$

In comparing the partial coefficient, $r_{13.4} = -0.1391$, with the corresponding gross correlation, $r_{13} = -0.4685$, it was found that there was little relation between the United States production, X_3, and the Minneapolis price, X_1, not already explained by world production, X_4. The world production was related to United States production ($r_{34} = +0.59$) and to the Minneapolis price ($r_{14} = -0.65$).

The calculation of these and other first-order partial correlation coefficients is shown in tabular form in the first part of table 3. The three gross coefficients used in the calculation of each first-order partial are given in the first two columns. The second term of the numerator, the product of the last two of each group of three gross coefficients, is given in column 3. In the calculation of $r_{12.3}$, this product was $(r_{13})(r_{23})$. The whole numerator, which was the difference between the first and second terms, is given in column 4. This value was $r_{12} - (r_{13})(r_{23})$.

The values of $\sqrt{1 - r_{13}^2}$ and $\sqrt{1 - r_{23}^2}$ are given in column 5. Their product, which was the denominator, is given in column 6. The first-order partial coefficients, $r_{12.3}$, etc., given in the last two columns, were determined by dividing the numerator (column 4) by the denominator.

It will be noted that the gross and partial coefficients have four decimal places, compared with three places in tables 1 and 2. The extra decimal place has no value except to secure greater accuracy in further calculations based on these coefficients.

The tabular determination of the first-order coefficients follows the algebraic formula

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

that of the second-order follows

$$r_{12.34} = \frac{r_{12.3} - (r_{14.3})(r_{24.3})}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}}$$

SECOND-ORDER COEFFICIENTS

The second-order partial correlation coefficients were obtained from various combinations of the first-order partial coefficients by the following formula:[14]

$$
\begin{aligned}
r_{12.34} &= \frac{r_{12.3} - (r_{14.3})(r_{24.3})}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \\
&= \frac{+0.7867 - (-0.5228)(-0.7691)}{\sqrt{1 - (-0.5228)^2}\sqrt{1 - (-0.7691)^2}} \\
&= \frac{+0.7867 - 0.4021}{(0.8525)(0.6391)} \\
&= \frac{0.3846}{0.5448} = +0.7059
\end{aligned}
$$

The partial coefficient, $r_{12.34} = +0.706$, is a measure of the relation between the Minneapolis and Liverpool prices, X_1 and X_2, after the effects of the United States and the world production, X_3 and X_4, have been eliminated.

The partial correlation between the Minneapolis price, X_1, and the United States production, X_3, with the effects of the Liverpool price, X_2, and the world production, X_4, eliminated, was $r_{13.24} = -0.565$ (table 3).

A general formula for the calculation of any order partial correlation coefficient for the X_1X_2 relationship from lower-order partials is:

$$
r_{12.34\cdots m} = \frac{r_{12.34\cdots(m-1)} - [r_{1m.34\cdots(m-1)}][r_{2m.34\cdots(m-1)}]}{\sqrt{1 - r_{1m.34\cdots(m-1)}^2}\sqrt{1 - r_{2m.34\cdots(m-1)}^2}}
$$

The formulas for $r_{13.24\cdots m}$, $r_{14.23\cdots m}$, etc., may be obtained by interchanging 3 and 2, 4 and 2, etc., in this formula.

INTERSERIAL CORRELATION

When the independent variables X_2 and X_3 are not only related to X_1, but are themselves related, the problem of interserial correlation is present. When the independent variables X_2 and X_3 are related, there is always the question whether the apparent effect of one independent variable is not merely a reflection of the other independent variable.

Factors affecting the price of corn may be considered as a concrete example of interserial correlation. During summers when pastures are

---

[14] The second-order partial, $r_{12.34}$, may also be obtained from the following partials:

$$
r_{12.34} = \frac{r_{12.4} - (r_{13.4})(r_{23.4})}{\sqrt{1 - r_{13.4}^2}\sqrt{1 - r_{23.4}^2}}
$$

poor, some farmers note that corn prices strengthen. The argument is that the resulting greater demand for feed raises the price of corn. Other farmers note that corn prices strengthen when the prospects for a corn crop are poor. The prospective short supply raises the price. It is apparent that both poor pastures and poor corn prospects make for high corn prices. Still other farmers observe that, in summers when corn prospects are poor, pastures also are frequently poor.

Those farmers who claimed that poor pastures raised corn prices failed to take into consideration that poor corn prospects occurred at the same time as poor pastures. Likewise, those who claimed that poor corn prospects raised prices failed to take poor pastures into account.

The two groups of farmers were in partial disagreement because of an interrelationship between the two factors that affected the price. In statistical jargon, this third relationship has been called interserial correlation.

In the problem of partial correlation, r_{12} and r_{13} might measure the relationship between: (*a*) the price of corn, X_1, and pasture, X_2; and (*b*) the price of corn, X_1, and the size of the crop, X_3. Then, r_{23} would be the interserial correlation between pasture and corn crop.

The problem of interserial correlation plays an important role in the analysis of three or more variables.

CALCULATION OF MULTIPLE COEFFICIENTS FROM PARTIALS

There are definite and well-known relationships among gross, partial, and multiple correlation coefficients. Formerly, instead of calculating partial coefficients from multiples, the common procedure was to determine multiples from the gross and partial correlations. The method was based on one gross coefficient and one partial of each order. To determine a multiple coefficient with three independent variables, one gross coefficient, one first-order and one second-order partial would be used in the following formula:[15]

$$R^2_{1.234} = 1 - (1 - r^2_{14})(1 - r^2_{13.4})(1 - r^2_{12.34})$$

The multiple coefficient for the Minneapolis price and the other three variables was $R_{1.234} = 0.846$.

$$\begin{aligned} R^2_{1.234} &= 1 - [1 - (-0.6494)^2][1 - (-0.1391)^2][1 - (+0.7059)^2] \\ &= 1 - (0.5783)(0.9807)(0.5017) \\ &= 1 - 0.2845 = 0.7155 \\ R_{1.234} &= \sqrt{0.7155} = 0.8459 \end{aligned}$$

---

[15] The subscripts 2, 3, and 4 are interchangeable. For example, this same multiple coefficient is also given by the formula

$$R^2_{1.234} = 1 - (1 - r^2_{12})(1 - r^2_{14.2})(1 - r^2_{13.24})$$

This agrees with the value of $R_{1.234}$ calculated through the solution of simultaneous equations.[16]

CHARACTERISTICS

The function of partial correlation analysis is the measurement of relationship between two factors, with the effects of one or more other factors eliminated. If the assumptions of the method are true for a series of data, the power of partial analysis is great. The problem of holding certain variables constant while the relationship between others is measured often presents itself in statistical analysis. ~~Partial~~ correlation is especially ~~useful in the analysis of~~ interrelated series. It is particularly pertinent to uncontrolled experiments of various kinds, in which such interrelationships usually exist. Most economic data ~~fall in~~ this category.

The problem of measuring partial relationship by tabulation methods is very difficult even when the number of observations is sufficient.

Partial analysis, like all correlation, has the advantage that the relationships are expressed concisely in a few well-defined coefficients.

It is adaptable to small amounts of data, and the reliability of the results can be rather easily tested.

The usefulness of partial analysis is somewhat limited by the following basic assumptions of the method:

The gross or zero-order correlations must have linear regressions.

The effects of the independent variables must be additively and not jointly related.

Because the reliability of the partial coefficient decreases as its order increases, the number of observations in gross correlations should be fairly large. Often the student carries the analysis beyond the limits of the data. This is a weakness of all research workers and to some extent can be guarded against by tests of reliability.

When the above assumptions have been satisfied, partial analysis still possesses the disadvantages of laborious calculations and difficult interpretation even for statisticians.

The interpretation of partial and multiple correlation results tends to assume that the independent variables have causal effects on the dependent variable. This assumption is sometimes true, but more often untrue in varying degrees. In describing the effects of the Liverpool price and United States and world production on the Minneapolis price of wheat, it was assumed that these effects were causal. There is nothing in the correlation method to prove whether the cause runs from the independent to the dependent variable, or vice versa. A person's knowledge and judgment must be his guide in deciding this point.

---

[16] Page 174.

For instance, for the Minneapolis and Liverpool prices of wheat, it was assumed that the Liverpool price was causal. However, the Minneapolis price may have had some small countereffect on the Liverpool price, and the correlation might represent more than the effect of the Liverpool price on the Minneapolis price.[17]

USES

Partial correlation is of greatest value when used in conjunction with gross and multiple correlation in the analysis of factors affecting variations in many kinds of phenomena.

The application of partial and multiple correlation to the relation of the supply of oats, supply of corn, and price of corn to the price of oats is summarized by the following coefficients:

November-April price of oats, Chicago, X_1	United States supply oats, X_2	$r_{12} = -0.77$	$r_{12.3} = -0.69$	$r_{12.34} = -0.86$	$R_{1.24} = 0.92$
		$r_{13} = -0.60$	$r_{12.4} = -0.80$	$r_{13.24} = +0.53$	$R_{1.234} = 0.94$
	United States supply corn, X_3	$r_{14} = +0.76$	$r_{13.2} = -0.44$	$r_{14.23} = +0.82$	
		$r_{23} = +0.46$	$r_{13.4} = +0.08$		
	Price corn at Chicago, X_4	$r_{24} = -0.83$	$r_{14.2} = +0.79$		
			$r_{14.3} = +0.58$		

The United States oat supply, X_2, the United States corn supply, X_3, and the purchasing power of the November-to-April price of corn at Chicago, X_4, explained 88 per cent of the variability in the price of oats, X_1 ($R^2_{1.234} = 0.88$). The high negative gross relation between oat price and supply ($r_{12} = -0.77$) and the high positive gross relation between oat and corn prices ($r_{14} = +0.76$) were improved slightly with the elimination of the effects of other factors ($r_{12.34} = -0.86$ and $r_{14.23} = +0.82$).

The relationship between the price of oats and supply of corn which first appeared to be negative ($r_{13} = -0.60$) became positive when the effects of the supply of oats and the price of corn were removed ($r_{13.24} = +0.53$). This is a good illustration of the complicated interrelationships that sometimes exist among independent variables. According to the partial correlation ($r_{13.24} = +0.53$), an increased production of corn called for a rise in the price of oats, instead of a decline as shown by the gross coefficient ($r_{13} = -0.60$). An increase in the supply of corn should have decreased the price of corn,[18] X_4, but the method holds the price of corn, X_4, constant. An increase in the supply of corn,[19] X_3, should also have been accompanied by an increase in the supply of oats, X_2. Since the supply of oats, X_2, was held constant, increasing the supply of corn, X_3, would be increasing the ratio of corn supply to oat supply. Such an increase would normally cause a decrease in the price of corn relative to the price of oats. Since the price of corn, X_4, was held constant and could not decrease, the ratio of the price of corn to the price of oats was decreased by raising the price of oats, X_1. Thus, an increase in the supply of corn, X_3, with the supply of oats, X_2, and the price of corn, X_4, held constant, resulted in an increase in the price of oats, X_1 ($r_{13.24} = +0.53$).

---

[17] In this connection, it is interesting to note that the Englishman says that the Liverpool price is made by Minneapolis or Chicago, while, in the United States, the general opinion is that the Minneapolis price is determined by the Liverpool price.

[18] $r_{34} = -0.83$.

[19] $r_{23} = +0.46$.

Although the gross coefficients indicated that the supply of corn had an important effect upon the price of oats ($r_{13} = -0.60$), the multiple coefficients showed that the supply of corn explained little of the variability in the price of oats not explained by the supply of oats and the price of corn[20] (compare $R^2_{1.234} = 0.89$ and $R^2_{1.24} = 0.85$).

OTHER MEASURES OF PARTIAL RELATIONSHIP

From time to time, the separate effects of the independent factors upon the dependent have been studied with measures other than partial correlation coefficients. One of these measures is the coefficient of part correlation given by the formula

$$_{12}r_{34} = \sqrt{\frac{b^2_{12.34}\sigma^2_2}{b^2_{12.34}\sigma^2_2 + \sigma^2_1(1 - R^2_{1.234})}}$$

If the multiple correlation coefficient, $R_{1.234}$, has been calculated, the part correlation can be easily obtained from the four known values, $R^2_{1.234}$, $b_{12.34}$, σ^2_2, and σ^2_1. The part correlation between Liverpool and Minneapolis prices of wheat was $_{12}r_{34} = +0.84$.

Like the partial correlation, $r_{12.34} = +0.71$, the part correlation $_{12}r_{34} = +0.84$ may be interpreted as a measure of the relationship existing between X_1 and X_2, with the effects of X_3 and X_4 eliminated.

---

[20] One of the limitations of the partial coefficient is that it is a relative measure rather than an absolute measure of the unexplained variability which is explained by the additional variable. For instance, when $R^2_{1.234}$ and $R^2_{1.24}$ were 0.89 and 0.85, the difference was 0.04 and the partial coefficient was $r_{13.24} = 0.53$. Relatively, the supply of corn explained 28 per cent of the variability in the price of oats not explained by the price of corn and supply of oats, $r^2_{13.24} = (0.53)^2 = 0.28$. Absolutely, X_3 explained only 4 per cent of the total variability in the price of oats in addition to that explained by the other two factors.

If the percentages of determination had been much smaller, for example, $R^2_{1.234} = 0.29$ and $R^2_{1.24} = 0.25$, again X_3 would have explained 4 per cent of the total variability in X_1 not already explained by X_2 and X_4. However, the proportion of the unaccounted-for variability explained would have been 5 per cent $\left(\frac{0.29 - 0.25}{1 - 0.25} = 0.053\right)$ rather than 28 per cent. The partial correlation coefficient would have been $r_{13.24} = 0.23$ instead of 0.53.

In part correlation, the effects of X_3 and X_4 are eliminated simultaneously with the consideration of X_2; in partial correlation, the effects of X_3 and X_4 are eliminated prior to the consideration of X_2.

The squared partial coefficient, $r^2_{12.34} = 0.50$, measured the proportion of the variability in X_1 not explained by X_3 and X_4 with no reference to X_2 in the linear relationship $X_1 = a + b_{13.4}X_3 + b_{14.3}X_4$, which was explained by the additional consideration of X_2 in the relationship

$$X_1 = a' + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

The squared part correlation coefficient, ${}_{12}r^2_{34} = 0.71$, measured the proportion of the variability in X_1 not explained by X_3 and X_4 considered simultaneously with X_2 in the relation $X_1 = a + b_{13.24}X_3 + b_{14.23}X_4$ which was explained by X_2 in the relation $X_1 = a' + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$. The variability in X_1 explained by X_2 in the simultaneous consideration of X_2, X_3, and X_4 was expressed as a proportion of the variability not explained in this consideration by X_3 and X_4 to get the part correlation. This variability in X_1 which was explained by X_2 may also be expressed as a proportion of the total variability in X_1. This proportion is measured by the squared beta coefficient given by the following formula:

$$\beta^2_{12.34} = b^2_{12.34}\frac{\sigma_2^2}{\sigma_1^2} \quad \text{or} \quad \beta_{12.34} = b_{12.34}\frac{\sigma_2}{\sigma_1}$$

The beta coefficient between Minneapolis and Liverpool prices of wheat was +0.83; and its square, 0.70. This indicated that, in the multiple relationship between the Minneapolis price and the three independent factors, the Liverpool price accounted for 70 per cent of the squared variability in the Minneapolis price.

The beta coefficient, $\beta_{12.34}$, like part correlation, is easily determined when the multiple coefficient, $R_{1.234}$, has been calculated.

The partial correlation, part correlation, and beta coefficients are measures of net relation between an independent and the dependent variables, but have somewhat different meanings and different values.

CHAPTER 12

CURVILINEAR CORRELATION

One of the assumptions underlying gross and multiple correlation analysis is that the relationships measured are linear. However, in many problems, the relationships do not follow the law of the straight line. For example, an increase in the amount of rainfall from scarcity to sufficiency may raise the yield of corn, while further increase beyond sufficient moisture may decrease yields. It has been found that many of the relationships between supply and price definitely follow the pattern of a curve. An increase in the production of cabbage from 70 to 80 per cent of normal resulted in a greater decline in the price of cabbage than that with an increase in production from 120 to 130 (figure 1).

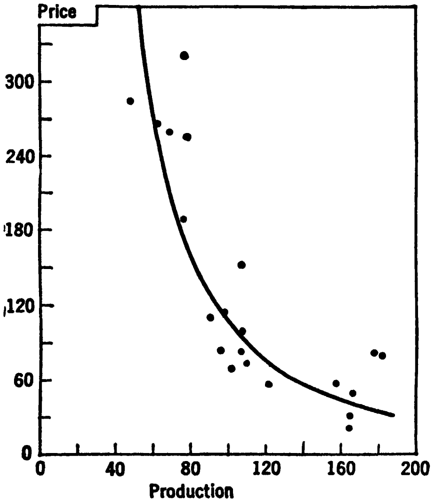


FIGURE 1.—RELATION OF THE PRODUCTION TO THE PRICE OF CABBAGE, UNITED STATES, 1920–1939

$\log Y = 5.6547 - 1.8150 \log X$

The average relationship is represented by the curve.[1]

INDEX OF CORRELATION

The research worker often needs a measure of relationship similar to the correlation coefficient, but one which takes into account the curvilinear nature of the relationship. A simple measure of this type is the index of correlation, ρ (rho). The index of correlation

---

[1] In determining the equation of this curve, the following simultaneous equations were solved for the values of a and b:

$$\Sigma \log Y = N \log a - b\Sigma \log X$$
$$\Sigma \log X \log Y = \log a\Sigma \log X - b\Sigma(\log X)^2$$

The values of a and b were substituted in the general equation

$$\log Y = \log a - b \log X$$

is similar to the correlation coefficient and may be written diagrammatically as follows:

$$\rho = \sqrt{1 - \frac{\text{Standard error squared about curve of relationship}}{\text{Standard deviation squared or variance}}}$$

and algebraically:

$$\rho = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}}$$

One of the formulas for the correlation coefficient is:

$$r = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}}$$

The only difference between the expressions for ρ and r is between the two standard errors of estimate, S_Y. Each squared standard error of estimate, S_Y^2, is the average of the squared residuals, $S_Y^2 = \Sigma z^2/N$. However, for ρ, the residual, z, is measured about a curve of relationship; while for r, the residual, z, is measured about a straight line fitted by the method of least squares. The close relationship between ρ and r may be pointed out by stating that r is a specialized type of ρ; that is, $\rho = r$ when the curve about which the residuals are measured is a least-squares straight line.

The meaning of ρ and ρ^2 in reference to the curve of relationship is substantially the same as the meaning of r and r^2 in reference to a straight line. The index of correlation squared, ρ^2, like r^2, is a measure of the proportion of the squared variability or variance in the dependent variable, Y, associated with differences in the independent variable, X. Rho, ρ, like r, is an abstract measure of the relationship[2] between the two variables considered.

In linear correlation, there is no problem of determining the pattern of relationship. A straight-line relationship is assumed, and the data are automatically fitted to a straight line in the calculation of r. In curvilinear correlation, the problem of determining the pattern of relationship is an important part of the analysis. A wide variety of mathematical or freehand curves can be used to show relationships. Some of these curves fit better than others.

The relationship between the production and price of cabbage was plotted, and a mathematical curve of the type $Y = \frac{a}{X^b}$ was fitted by the method of least squares (figure 1). The curve of relationship was found to be $Y = \frac{451,500}{X^{1.8150}}$, or $\log Y = 5.6547 - 1.8150 \log X$. The values of Y estimated from this curve of relationship between production and price were obtained by substituting in the equation the actual production, X, and solving for the estimated price, Y'. In 1920, cabbage production was high, $X = 178$, and the price estimated by the curve was low, $Y' = 37$ (table 1, middle section). This was determined as follows:

---

[2] A detailed explanation of the meaning of r^2 and r is given on pages 143 to 146.

$$\begin{aligned} \log Y' &= 5.6547 - 1.8150(\log 178) \\ &= 5.6547 - 1.8150(2.250) \\ &= 5.6547 - 4.0838 = 1.5709 \\ Y' &= 37 \end{aligned}$$

The estimated price for the other years was determined in the same way. The residual for 1920 was obtained by subtracting the estimated price, Y', from the actual price, Y.

$$z = Y - Y' = 82 - 37 = 45$$

The next step consisted of squaring the residuals and summing the squares. From the sum of the squared residuals, $z^2 = 55,745$, the standard error of estimate squared was easily derived by dividing by the number of observations, $N = 20$, as follows:

$$S_Y^2 = \frac{\Sigma z^2}{N} = \frac{55,745}{20} = 2,787$$

The squared standard deviation[3] in the actual price was 8,426. The index of correlation was as follows:

$$\rho_{\left(Y - \frac{a}{X^b}\right)} = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}} = \sqrt{1 - \frac{2,787}{8,426}} = \sqrt{1 - 0.331} = \sqrt{0.669} = 0.818$$

This coefficient, 0.818, is a measure of the degree of relationship between the production and price of cabbage. Rho squared, 0.669, is the proportion of the squared variability or variance in the price of cabbage which can be explained by production. These values of ρ and ρ^2 are peculiar to the particular curve used to express the relationship and

---

[3] Footnote to table 1.

TABLE 1.—CALCULATION OF RESIDUALS FROM CURVES AND INDEXES OF CORRELATION

Deviations of Actual Price of Cabbage from the Price Estimated from Mathematically Determined and Freehand Curves, 1920–1939

Year	Original data*			Mathematical curve $\log Y = 5.6547 - 1.8150 \log X$			Freehand curve		
	Index, production X	Index, price Y	Y^2	Estimated price Y'	Residuals $(Y - Y')$ z	Residuals squared z^2	Estimated price Y'	Residuals $(Y - Y')$ z	Residuals squared z^2
1920	178	82	6,724	37	45	2,025	35	47	2,209
1921	47	285	81,225	417	– 132	17,424	480	– 195	38,025
1922	164	23	529	43	– 20	400	37	– 14	196
1923	76	189	35,721	174	15	225	177	12	144
1924	157	58	3,364	47	11	121	39	19	361
1925	93	153	23,409	121	32	1,024	108	45	2,025
1926	106	98	9,604	95	3	9	81	17	289
1927	109	74	5,476	91	– 17	289	77	– 3	9
1928	68	259	67,081	213	46	2,116	246	13	169
1929	96	84	7,056	114	– 30	900	102	– 18	324
1930	107	83	6,889	94	– 11	121	79	4	16
1931	97	114	12,996	112	2	4	98	16	256
1932	121	58	3,364	75	– 17	289	62	– 4	16
1933	76	321	103,041	174	147	21,609	177	144	20,736
1934	165	31	961	43	– 12	144	36	– 5	25
1935	90	110	12,100	128	– 18	324	119	– 9	81
1936	77	256	65,536	170	86	7,396	172	84	7,056
1937	101	69	4,761	104	– 35	1,225	90	– 21	441
1938	166	50	2,500	42	8	64	36	14	196
1939	61	266	70,756	260	6	36	325	– 59	3,481
Total		2,663	523,093	—	+109	55,745	—	+87	76,055
Average		133.15	26,154.65						

$$S_Y^2 = \frac{55{,}745}{20} = 2{,}787 \qquad \rho_{(Y=a/X^b)} = \sqrt{1 - \frac{2{,}787}{8{,}426}} = 0.818$$

$$S_Y^2 = \frac{76{,}055}{20} = 3{,}803 \qquad \rho_{(\text{freehand})} = \sqrt{1 - \frac{3.803}{8{,}426}} = 0.741$$

* The column Y^2 is irrelevant except that it is used in the calculation of the squared standard deviation, as follows:

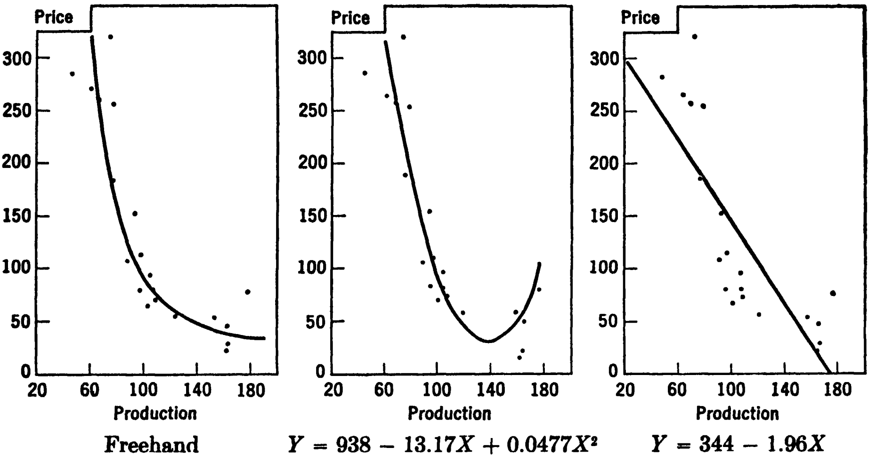
$$\sigma_Y^2 = AY^2 - (AY)^2 = 26{,}154.65 - (133.15)^2 = 8{,}425.73$$

should be written $\rho_{\left(Y=\frac{a}{X^b}\right)}$ and $\rho^2_{\left(Y=\frac{a}{X^b}\right)}$. If some other curve were used, or if this curve were fitted by some other method, the value of ρ would be somewhat different.

RHO FROM DIFFERENT CURVES

The authors approximated the relationship between the production and price of cabbage with a freehand curve (figure 2, left). The estimated price of cabbage was read from the plotted freehand curve. These estimated prices, Y', were recorded in the right-hand side of table 1. The residuals about the freehand curve were obtained, squared, and summed. The squared standard error of estimate was $S_Y^2 = \frac{\Sigma z^2}{N} = \frac{76,055}{20} = 3,803$. The index of correlation was as follows:

$$\rho_{(\text{freehand})} = \sqrt{1 - \frac{S_Y^2}{\sigma_Y^2}} = \sqrt{1 - \frac{3,803}{8,426}} = \sqrt{1 - 0.451} = \sqrt{0.549} = 0.741$$



Freehand

$Y = 938 - 13.17X + 0.0477X^2$

$Y = 344 - 1.96X$

FIGURE 2.—THE RELATION BETWEEN THE PRODUCTION AND PRICE OF CABBAGE DESCRIBED BY A FREEHAND CURVE, A PARABOLA, AND A STRAIGHT LINE

The freehand approximation was a more accurate description of the average relationship between the production and price of cabbage than the parabola[4] or the straight line.[5]

The relationship between production and price as measured by the freehand curve, $\rho_{(\text{freehand})} = 0.741$, was not so high as that measured by the curve fitted mathematically, $\rho_{(Y-a/X^b)} = 0.818$ (table 1). The general shapes of the two curves were the same, but the freehand curve fitted the data less accurately. This was especially true for the very short crop of 1921. Although both curves overestimated the price, the residual squared for the freehand curve was more than twice that for the mathematical curve (38,025 and 17,424, respectively, table 1).

---

[4] The parabola was fitted by the averages method. The average supply and the average price were calculated for the 6 years of lowest production, 8 years of average production, and 6 years of large production. Substituting these averages for X and Y in the general equation $Y = a + bX + cX^2$, there were three simultaneous equations which were solved for a, b, and c. These values were substituted in the general equation to obtain the average equation $Y = 938 - 13.17X + 0.0477X^2$.

[5] The straight line was calculated by the usual correlation methods.

When the parabola, $Y = a + bX + cX^2$, was fitted to the production and price of cabbage (figure 2, center), the index of correlation was

$$\rho_{(Y=a+bX+cX^2)} = \sqrt{1 - \frac{2{,}167}{8{,}426}} = \sqrt{1 - 0.257} = \sqrt{0.743} = 0.862$$

When the straight line, $Y = a + bX$, was used (figure 2, right), the coefficient of correlation was

$$r_{YX} = \sqrt{1 - \frac{2{,}995}{8{,}426}} = \sqrt{1 - 0.355} = \sqrt{0.645} = -0.803$$

The results of the four attempts to establish the relationship between the production and price of cabbage were as follows:

$$\rho_{(Y=a/X^b)} = 0.82 \qquad \rho_{(Y=a+bX+cX^2)} = 0.86$$
$$\rho_{(\text{freehand curve})} = 0.74 \qquad \rho_{(Y=a+bX)} = r_{YX} = -0.80$$

All the indexes of correlation were high, indicating the close relationship between the production and price. The authors assumed that a curve of the type $Y = a/X^b$ or the freehand curve most accurately described the law of relationship between the production and price of cabbage. However, indexes of correlation based on these curves, 0.82 and 0.74, were about the same as or lower than those based on a straight line or parabola whose mathematical laws did not conform to the law of relationships assumed to exist between the production and price of cabbage. This is a good illustration of the fact that an illogical curve often gives a better fit than a logical curve. The parabola gives the closest mathematical fit but does not show the correct principle of the relationship.

EFFECT OF EXTREME RESIDUALS

There were several reasons why the largest values of ρ were obtained from the curves whose mathematical laws did not conform with the data. One of the most important was chance. With only 20 years of data and a high degree of variability, the probability would be high that the inclusion of one unusual year would result in a higher ρ from an illogical curve than from a logical one.

One or two very large residuals may be extremely important in determining the values of r or ρ. If the year 1921 were not considered, ρ would probably have been higher from the freehand curve than from the straight line. Of the total squared residuals about the freehand curve, $\Sigma z^2 = 76{,}055$, one-half, 38,025, were contributed by one year, 1921 (table 1). If it is assumed that the residual from the freehand curve for 1921 was zero, the sum of the squared residuals would have been only 38,030, instead of 76,055. The index of correlation, $\rho_{(\text{freehand})}$, would have been 0.88 instead of 0.74.

Much of the effect of the 1921 residual on $\rho_{(\text{freehand})}$ may be traced to the squaring process. The squared residual for 1921 was 50 per cent of the total. The actual residual, without respect to sign, was only 26 per cent of the total.

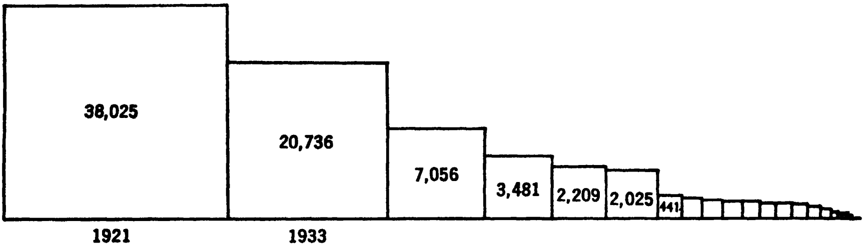


FIGURE 3.—SQUARED RESIDUALS ABOUT A FREEHAND CURVE ARRANGED BY SIZE

Production and Price of Cabbage

The height of each square represents the size of the residual; and the area, the squared residual. The very large squared residual for 1921 accounted for 50 per cent of the total of the squared residuals; and the two very large residuals, 1921 and 1933, accounted for 77 per cent of the total.

The sum of 14 small residuals to the right represents only 3 per cent of Σz^2.

The squared residuals for the freehand curve in table 1 were arranged by size and shown graphically in figure 3. Over half the total squared residuals were included in the two years of greatest deviation from the freehand curve. About two-thirds of the squared residuals comprised a small proportion of the total and were not important in reducing ρ.

EFFECT OF FLEXIBILITY OF CURVES

One of the reasons that $\rho_{(Y=a+bX+cX^2)}$ was larger than $\rho_{(Y=a/X^b)}$ and $\rho_{(Y=a+bX)}$ was the greater flexibility of the curve on which it was based. The parabola, $Y = a + bX + cX^2$, has three constants, a, b, and c, which make it more flexible than the other two curves with only two constants, a and b. Obviously, the freehand curve is most flexible of all. The index of correlation, $\rho_{(\text{freehand})} = 0.74$, was the lowest, not because the curve lacked flexibility but because it was drawn least accurately.

EFFECT OF METHOD OF FITTING CURVES

The curve $Y = \frac{a}{X^b}$ was fitted by the methods of least squares, selected points,[6] and semi-averages, and approximated by the freehand method. The respective indexes of correlation were:

$$\rho_{LS} = 0.82, \quad \rho_A = 0.52, \quad \rho_{SP} = 0.32, \quad \rho_{FH} = 0.74$$

[6] The selected-points and semi-averages methods are described on page 78.

There was more variation due to the method of fitting than to the type of curve used. In general, the ρ_{LS} should be the largest; and the ρ_{SP}, the smallest. For less curvilinear data, the differences would not always be so large as in this example.

TABLE 2.—EFFECT OF MEASURING RESIDUALS IN NATURAL NUMBERS AND IN LOGARITHMS*

PRODUCTION AND PRICE OF CABBAGE, 1920–1939

Year	Original data		log X	log Y	Calculation of residuals in logarithms			Calculation of residuals in natural numbers†		
	Index, production X	Index, price Y			Estimated logarithms log Y'	Residuals log Y − log Y' z	Residuals squared z^2	Estimated price Y'	Residuals $Y - Y'$ z	Residuals squared z^2
1920	178	82	2.2504	1.9138	1.5702	0.3436	0.1181	37	45	2,025
1921	47	285	1.6721	2.4548	2.6198	−0.1650	0.0272	417	−132	17,424
1922	164	23	2.2148	1.3617	1.6348	−0.2731	0.0746	43	− 20	400
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
1937	101	69	2.0043	1.8388	2.0169	−0.1781	0.0317	104	− 35	1,225
1938	166	50	2.2201	1.6990	1.6252	0.0738	0.0054	42	8	64
1939	61	266	1.7853	2.4249	2.4144	0.0105	0.0001	260	6	36
Total	—	2,663	—	40.3025	40 2982	—	0.4513	2,554	—	55,745
Average	—	133.15	—	2.0151	2.0149	—	0.0226	127.7	—	2,787.25

* The columns for the calculation of $\sigma^2_{\log Y}$ are not shown. $\sigma^2_{\log Y} = A(\log Y)^2 - (A \log Y)^2$, or

$$\sigma^2_{\log Y} = 4.1619 - (2.0151)^2 = 4.1619 - 4.0606 = 0.1013$$

$$\rho_{\log Y} = \sqrt{1 - \frac{S^2_{\log Y}}{\sigma^2_{\log Y}}} = \sqrt{1 - \frac{0.0226}{0.1013}} = \sqrt{1 - 0.2231} = \sqrt{0.7769} = 0.881$$

† From table 1. The index of correlation was $\rho = 0.818$.

EFFECT OF METHOD OF MEASURING RESIDUALS

The size of rho is also influenced by the method used in measuring residuals. Residuals and standard deviations can be expressed in terms of natural numbers, logarithms, reciprocals, or the like. The residual, z, might be $Y - Y'$, $\log Y - \log Y'$, or $\frac{1}{Y} - \frac{1}{Y'}$.

In the problem on the production and price of cabbage, the residuals were expressed in logarithms ($\log Y - \log Y'$),[7] and the resulting index of correlation was $\rho = 0.881$ (table 2). The results obtained with natural numbers are given in the last three columns of table 2. The index of correlation was slightly less, $\rho = 0.818$. These indexes of correlation were based on the following residuals:

---

[7] For the curve $Y = \frac{a}{X^b}$ used in this example, there is more justification in expressing residuals in logarithms than in natural numbers because the curve was fitted by specifying that $\Sigma (\log Y - \log Y')^2$ be a minimum. The authors measured residuals in natural numbers in table 1 merely to simplify the explanation of rho.

$$\Sigma z^2_{(\log Y')} = 0.4513 \text{ (logarithms)}$$

and

$$\Sigma z^2_{(Y')} = 55.745 \text{ (natural numbers)}$$

To express a residual as the difference between two logarithms is the same as expressing the difference in natural numbers as a ratio of the actual to the estimated price. The effect of this method is to minimize the importance of residuals in the short-crop years. The greatest residuals happened to be in the two small-crop years. If they had been in two large-crop years, rho based on logarithm residuals might have been less than rho based on natural residuals.

When curves are fitted by the least-squares method, rho is usually largest when the residuals are measured in the same terms used in deriving the normal equations for the curve. For example, the normal equations for the curve $Y = \frac{a}{X^b}$ were derived by specifying that $\Sigma(\log Y - \log Y')^2$ be a minimum. Therefore, rho based on residuals $(\log Y - \log Y')$ would in a majority of cases be larger than rho based on $(Y - Y')$ or $\left(\frac{1}{Y} - \frac{1}{Y'}\right)$.

IMPORTANCE OF DEFINING RHO

The index of correlation for any given series of data will differ with the different types of curves used, with the amount of flexibility in the curves, with the method of fitting the curves, and with the method of expressing the residuals.

The values of ρ calculated from the production and price of cabbage were:

$$\rho_{(Y=a/X^b)(LS)(\text{natural numbers})} = 0.82$$

$$\rho_{(Y=a/X^b)(LS)(\text{logarithms})} = 0.88$$

$$\rho_{(Y=a/X^b)(A)(\text{natural numbers})} = 0.52$$

$$\rho_{(Y=a/X^b)(SP)(\text{natural numbers})} = 0.32$$

$$\rho_{(Y=a/X^b \text{ approximation})(FH)(\text{natural numbers})} = 0.74$$

$$\rho_{(Y=a+bX+cX^2)(A)(\text{natural numbers})} = 0.86$$

$$\rho_{(Y=a+bX)(LS)(\text{natural numbers})} = r_{YX} = -0.80$$

One can obtain only one value of r from a given series, but as many different values of ρ as there are kinds of curves, methods of fitting, and methods of expressing the residuals. The value of ρ is meaningless unless the above conditions are indicated.

PROBLEMS IN CHOOSING RHO

Because so many different values of ρ can be obtained from a given series of data, the student is usually in a quandary as to which ρ to use. The question simmers down to (*a*) choice of the curve of relationship, (*b*) the methods of fitting the curve, and (*c*) measuring the residuals.

The choice of curve should be based both on the data and on the expected law of relationship. A measure of the goodness of fit of a curve of relationship is ρ itself. However, ρ about a highly flexible curve will usually be higher than about simple curves with few constants such as $Y = \frac{a}{X^b}$, $Y = a + bX$, and the like. The student should be guided not only by the size of ρ but also by the logic of the curve. $\rho_{(Y=a+bX+cX^2)} = 0.86$ was higher than $\rho_{(Y=a/X^b)(\text{natural numbers})} = 0.82$. On the basis of the size of ρ, the first curve would be preferable, but, on the basis of logic, the second should be used.[8] $\rho_{(Y=a/X^b)(\text{natural numbers})} = 0.82$ was about the same as $\rho_{(Y=a+bX)} = 0.80$. Since both curves have two constants, a and b, they have the same degree of flexibility. The straight line is the simpler curve, but the curve $Y = \frac{a}{X^b}$ was preferable because its mathematical laws approximated more closely the expected law of relationship between the production and price of cabbage.

When the law of relationship is not known, the student should, in general, use the curve that fits most closely. However, it is always advisable to be conservative and choose curves with few degrees of flexibility. If a person is not conscientious, he may obtain ρ of any size he desires, depending on the flexibility, that is, the number of constants in the curve. For example, a freehand curve could have been drawn through each of the 20 points of the cabbage data, and then $\Sigma z^2 = 0$ and $\rho = 1.00$. Similarly, if the mathematical curve had 20 constants, $Y = a + bX + cX^2 + \cdots + rX^{17} + sX^{18} + tX^{19}$, then $\Sigma z^2 = 0$ and $\rho = 1.00$. Obviously, there is no justification for such highly flexible curves. The relationships shown by such curves and high indexes of correlation based on them are unreliable. The flexibility of curves should be limited to one or two bends, or to two or three constants. Unless the data depart considerably from linearity, the straight line is usually the safest.

In choosing a method of fitting a given curve, the student should be guided by closeness of fit and ease of calculation. The highest values of ρ result from the closest fits, and the closest fits usually result from the least-squares method. However, it is often possible to approximate these curves by freehand or other methods sufficiently closely to obtain values of ρ approaching those for least-squares curves. From the standpoint of ease of calculation, the best method is the freehand; and the most laborious, the least-squares.

---

[8] For large crops, the curve $Y = a + bX + cX^2$ gave estimated prices higher than for average crops.

The problem of choosing a method of measuring residuals hinges on the method of fitting the curve. When the least-squares method is used, in general, the residuals should be in the same terms used in the normal equations. When the curve is fitted by some method other than least squares, the student may express residuals in natural numbers or logarithms. If he desires a residual easy to calculate and easy to understand, he should use natural numbers. If he desires a residual which expresses the difference as a percentage, he should use logarithms.

CHARACTERISTICS OF RHO

The index of correlation, ρ, like the coefficient of correlation, r, is a measure of the association between two variables.

The squared index, ρ^2, like r^2, is a proportionate measure of the squared variability or variance in one factor associated with differences in the other.

Although $r_{YX} = r_{XY}$, it is not true that $\rho_{YX} = \rho_{XY}$. Even when the curve is fitted by least squares, the exact size of ρ depends upon which variable is considered dependent.

The curves on which ρ is based are frequently more valuable than ρ itself.

The index of correlation has several advantages over other methods of analyzing association:

1. The index of correlation is a concise measure of the degree of relationship. The curve on which it is based is a concise description of the nature of the relationship.
2. The *index* of correlation, unlike the *coefficient* of correlation, takes into account the curvilinear nature of the association. Since any type of curve can be used as a basis for ρ, the method is quite flexible.
3. Rho from freehand curves is somewhat easier to calculate than r, which is always based on a least-squares straight line.
4. With the index of correlation, average relationships can be obtained from a smaller amount of data than that necessary for methods based on the comparison of averages.

The index of correlation has important disadvantages:

1. Because the method is flexible and unstandardized, many different types of curves and different values of ρ can be obtained from the same data.
2. Because many different curves are possible, the reliability of any one curve and of its ρ is questionable. Flexibility, though desirable in

showing the exact nature of the relationship, decreases the reliability of the analysis. Too often the reliability of ρ and its curve is overrated.

USES

The primary use of the index of correlation is to show the nature and degree of association or relationship. To many people, the nature of the relationship between production and price of cabbage would be evident from the curve fitted to the data (figure 1, page 200). As the size of the crop increased, the price decreased at a decreasing rate. To others, the relationship would be more understandable in tabular form. The price of cabbage for productions of 50, 90, and so on can be read directly from the curve and set forth as follows:

PRODUCTION	PRICE
50	372
90	128
130	66
170	40

The extent to which the association shown by the curve or the above table was true was given by $\rho_{(Y=a/X^b)} = 0.82$. Rho, ρ, is a measure of the degree of relation, but is meaningless to most persons. It is not so valuable as the curve or table which shows the nature of the relationship.

Fish[9] used the index of correlation to measure the relationship between prices paid for all foods and prices paid for individual foods. When prices paid for tomatoes were compared with those for all foods, the index of correlation was calculated in terms of several curves, and two were published:

$$\rho_{(\text{log curve})} = 0.603 \qquad \rho_{(\text{parabola})} = 0.609$$

With each increase in the price of all foods purchased, prices paid for tomatoes rose sharply at a slightly decreasing rate.

Retail prices of beans, lettuce, fish, beef, pork, and flour increased at a decreasing rate. Prices of coffee, cheese, bacon, salmon, pineapples, and peaches increased at an increasing rate.

Bennett[10] used ρ in studying the price of oats in relation to the price of corn from 1897 to 1938. He found that, when the oat supply was high relative to the corn supply, the price of oats was low relative to corn. The index of correlation was $\rho_{(\text{freehand})} = 0.83$. When the supply of oats was about 70 per cent of the corn supply, the price of oats was 139 per cent of the price of corn. When the supply of oats was high, 130, the price was low, 85.

---

[9] Fish, M., Buying for the Household, Cornell University Agricultural Experiment Station Bulletin 561, pp. 40–45, June 1933.

[10] Bennett, K. R., The Price of Feed, unpublished manuscript, Cornell University, 1940.

CHAPTER 13

INDEX OF MULTIPLE CORRELATION

The coefficient of multiple correlation is based on the linear relationships which exist between the dependent and each independent variable. These relationships are combined into a multiple relation expressed as follows:

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

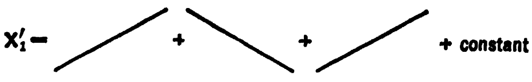
The multiple correlation coefficient, like all correlation coefficients and indexes, is given by the expression $\sqrt{1 - \frac{S^2}{\sigma^2}}$. However, S^2 is the average of the squared residuals of the actual X_1 from X_1' estimated from the linear relationship given above.

The index of multiple correlation is based on the curvilinear relationships which exist between the dependent and each independent variable. These relationships are also combined into a multiple relation which might be expressed in a general equation, as follows:

$$X_1 = a + f_2(X_2) + f_3(X_3) + f_4(X_4)$$

Likewise, the index of multiple correlation,[1] ρ, is given by the expression $\sqrt{1 - \frac{S^2}{\sigma^2}}$. However, X_1' is estimated from curvilinear relationships here. The only difference between R and ρ is in the value of S^2.

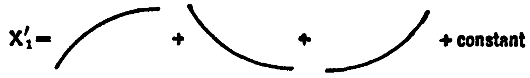
The coefficients, R and ρ, are also similar in that the effects of the independent variables are assumed to be additive. In each case, the estimated value of the dependent variable, X_1', is a sum of several single estimates. For example, in linear analysis, X_1' is the resultant of three linear relations, which might be as follows:



---

[1] The same symbol, ρ, is used to indicate the curvilinear correlation between one independent and the dependent variable and also between several independent and the dependent variable.

or in curvilinear analysis, as follows:



In linear analysis, the multiple relationship was determined mechanically and mathematically by the method of least squares. Some of the more simple curvilinear multiple relationships can also be determined by the method of least squares. To measure those relationships which are not mathematically simple and to explore the nature of relationships about which nothing is previously known, the method of least squares is practically useless. However, efficient methods of approximating those relationships have been devised.

LEAST-SQUARES ANALYSIS

One can use the least-squares method for curvilinear analysis by converting independent variables from natural numbers to curvilinear functions, such as logarithms, squares, reciprocals, and the like. A multiple correlation coefficient may be determined from such converted values. For example, if a multiple relation is of the type

$$X_1 = b_{12.34} \log X_2 + b_{13.24}X_3^2 + b_{14.23}\sqrt{X_4}$$

X_2 is converted to the logarithm of X_2; X_3, to X_3^2; and X_4, to $\sqrt{X_4}$. The coefficients of regression and correlation are then determined by the usual methods.[2] To use this method, the student must know in advance the general nature of the relation between X_2 and X_1, X_3 and X_1, and X_4 and X_1.

The linear multiple correlation coefficient for the acreage of corn in North Carolina, X_1, and the United States farm price of corn and cotton the preceding year, X_2 and X_3, and the stocks of corn on North Carolina farms, X_4, was $R_{1.234} = 0.666$ (table 1). The equation of relationship was

$$X_1 = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$$

indicating that the acreage of corn increased after high corn prices, decreased after high cotton prices, and increased when stocks of corn on farms were large.

The above equation assumed that each relationship was linear; other possibilities were ignored. When plotted on graph paper, however, the relations between X_1 and X_2 and between X_1 and X_3 appeared to be curvilinear. The pattern of the X_1X_2 relationship seemed to resemble a curve increasing at a decreasing rate. The X_1X_3 relationship seemed to resemble a curve decreasing at an increasing rate. The X_1X_4 relationship seemed to increase at a constant rate. After studying various curves, it appeared that the three relationships could be described as follows:

---

[2] Pages 168 to 176.

X_1X_2, log curve	$X_1 = a + b \log X_2$
X_1X_3, second degree polynomial	$X_1 = a + \cancel{(bX_3)} + cX_3^2$
X_1X_4, straight line	$X_1 = a + bX_4$

The equation for the multiple relationship was thought to be

$$X_1 = a + b_{12.34} \log X_2 + b_{13.24}X_3^2 + b_{14.23}X_4$$

This is obviously not a linear equation in X_2 or X_3, but it is linear in the logarithm of X_2, in the square of X_3, and in the natural values of X_4.

TABLE 1.—VARIABLES FOR LINEAR AND CURVILINEAR MULTIPLE CORRELATION ANALYSIS

ACRES OF CORN IN NORTH CAROLINA, X_1; AND THE UNITED STATES FARM PRICE OF CORN THE PRECEDING YEAR, X_2; THE UNITED STATES FARM PRICE OF COTTON THE PRECEDING YEAR, X_3; AND STOCKS OF CORN ON NORTH CAROLINA FARMS, X_4

Year	Data for linear analysis				Data for curvilinear analysis			
	X_2	X_3	X_4	X_1	$\log X_2$	X_3^2	X_4	X_1
1	72	8	11	27	1.86	64	11	27
2	45	9	16	26	1.65	81	16	26
3	50	14	17	27	1.70	196	17	27
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
23	47	20	29	23	1.67	400	29	23
24	61	14	18	24	1.79	196	18	24
25	42	11	19	24	1.62	121	19	24
Total	1,340	298	574	675	42.99	3,770	574	675
Average	53.6	11.92	22.96	27.0	1.720	150.8	22.96	27.0

$\sigma_1^2 = 3.60$ $R_{1.234} = 0.666$

$X_1 = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$

$\rho = 0.701$

$X_1 = 10.02 \log X_2 - 0.01042X_3^2 + 0.1091X_4 + 8.836$

In linear multiple correlation analysis, the four variables were in their natural forms. In this curvilinear analysis, the variable X_2 was converted to its logarithm; and X_3, to its square. For the first year, $X_2 = 72$. The value used in the curvilinear problem was 1.86, the logarithm of 72. The value of X_3, which was 8, was converted to its square, 64. For the second year, $\log X_2 = 1.65$ was used instead of $X_2 = 45$; and $X_3 = 9$ was discarded for $X_3^2 = 81$ (table 1).

With the natural forms of X_4 and X_1 and the new forms of X_2 and X_3, the usual procedure was followed in determining the multiple regression and correlation coefficients. The index of multiple correlation[3] was 0.701, which was somewhat, though not greatly, higher than $R = 0.666$. The new multiple relationship measured by $\rho = 0.701$ was a combination of two curves and a straight line. The curves apparently fit a little better than the two straight lines they replaced. The fact that these two relationships appeared definitely curvilinear in advance might lead one to expect an increase from R to ρ greater than 0.035 ($0.701 - 0.666 = 0.035$). Perhaps the authors used the wrong functions of X_2 and X_3. It might be that no set of curves as simple as these will fit the data more closely.[4]

The detailed nature of the multiple curvilinear relationship is given by

$$X_1 = 10.02 \log X_2 - 0.01042X_3^2 + 0.1091X_4 + 8.836$$

Apparently, the acreage of corn increased after high corn prices, decreased after high cotton prices, and increased slightly with large stocks. The directions of these and of the corresponding linear relationships were the same. However, the patterns of the relationships were necessarily different (figure 1). They were stipulated in one case as all linear; and, in the other case, as one logarithmic, one geometric, and one linear. Stated more simply: as the price of corn rose, the acreage planted the year following also rose, but at a decreasing rate. As the price of cotton rose, the acreage of corn decreased at an increasing rate. The effects on the acreage were greatest when the price of corn was low or when the price of cotton was high.[5]

---

[3] Some workers call this measure R when it is determined by this procedure. It is true that ρ in this case is really R based on $\log X_2$, X_3^2, and X_4. In another sense, it is ρ based on X_2, X_3, and X_4.

[4] Those who expected ρ to increase above R because of the increased flexibility obtained in switching from straight lines to curves did not realize that neither the logarithm nor the square of a number is any more flexible than the number itself. Any increased flexibility in ρ is due to the wide range of curves on which it can be based. Rho may have as many values as there are types of curves, whereas R always has only one value.

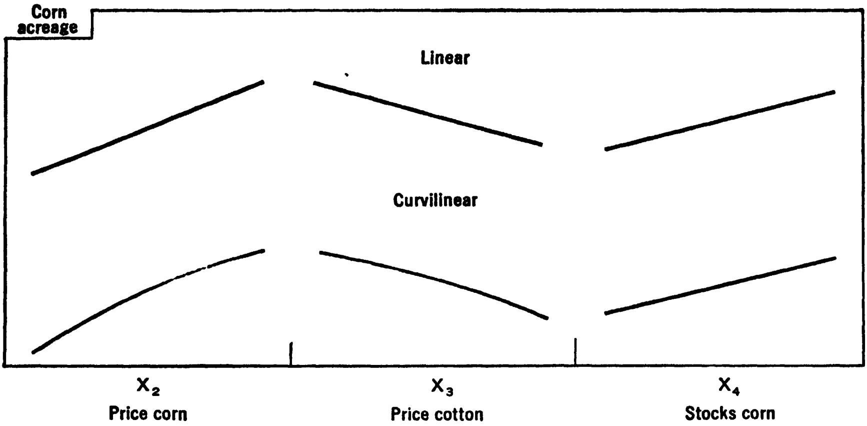


FIGURE 1.—LINEAR AND CURVILINEAR RELATIONSHIPS[6]

ACREAGE OF CORN AND PRICES OF CORN AND COTTON, AND STOCKS OF CORN, NORTH CAROLINA

Least-Squares Analysis

The linear relationships (above) assumed constant changes in corn acreage with unit changes in prices of corn and cotton.

The curvilinear relationships (below) assumed that changes in the price of corn were most effective upon acreage when the price was low, and that changes in cotton prices were most effective when the price was high.

The curvilinear assumptions were logical, and the curves fit the data a little better than the straight lines.

The results of this problem certainly need not be taken by the student as final. The authors obtained a higher ρ from the curvilinear than from the linear relationships, but there is no reason to believe that ρ might not have been greater, had more nearly correct functions of X_2 and X_3 been employed. For example, the authors suspect that X_3^3 should have been used, rather than X_3^2. This suggestion or others might be tried in another attempt to improve the relationship. Perhaps the trouble lay in insufficient flexibility in the curves. Perhaps X_3^2 should have been changed to $(X_3 - c)^2$, or the equivalent $(X_3^2 - 2cX_3 + c^2)$.

---

[5] Apparently, the acreage of corn, X_1, increased with the stocks of corn on farms the preceding March, X_4. This does not appear logical. Since the relationship was not very close, it may have been due to chance. It may also have been due to a failure to eliminate the slight secular trends in acreage and stocks of corn. If any real trends did exist, they would probably have been in the same direction for acreage and stocks of corn because the stocks depend on production which would be influenced by acreage.

[6] Linear, $X_1 = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$.
Curvilinear, $X_1 = 10.02 \log X_2 - 0.01042X_3 + 0.1091X_4 + 8.836$.

At any rate, the student, in seeking some improvement in the accuracy of the curves or in the size of ρ, would have no method of foretelling in advance *exactly* which functions of the independent variables to use. In his attempts he may waste considerable time in (1) trying a large number of curves, and (2) trying more degrees of flexibility. He may even conclude that the correct function is one such as

$$X_1 = a + b \log (X_2 + c)$$

where c is neither known in advance nor can be determined by the method of least squares. Thus, the least-squares method of determining ρ may be not only cumbersome, but even impossible.

APPROXIMATION ANALYSIS

Where the exact nature of multiple relationships is not known in advance, the least-squares method of curvilinear analysis is often a long story of many unsuccessful trials with different kinds of curves. Moreover, least-squares analysis is limited to the types of curves which can be fitted by this method. Because of these two difficulties, approximation methods, which are more flexible and more efficient for curvilinear analysis, were devised. Two approximation methods will be discussed: (*a*) graphic method from linear multiple regression, and (*b*) short-cut graphic method.

Approximation from Linear Multiple Regression

First Approximations

Ezekiel[7] developed a method of approximating curvilinear relationships from linear net regressions. The first step in this method is to determine the multiple correlation coefficient, $R_{1.234}$, the net regression coefficients, $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$, and the net regression equation, $X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$. In the corn acreage problem, $R_{1.234} = 0.666$, the linear relationship was $X_1 = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$ (table 1).

The second step consists in determining the residuals, z, for each year. The residual is merely the difference between the actual X_1 and the estimated X_1', based on the multiple regression. Each estimated X_1' is obtained by substituting the corresponding values of X_2, X_3, and X_4 in the multiple regression equation. For example, for the first year,

$$X_1' = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$$
$$X_1' = 0.0842(72) - 0.2312(8) + 0.0971(11) + 23.01$$
$$X_1' = 6.1 - 1.8 + 1.1 + 23.0 = 28.4$$

---

[7] Ezekiel, M., A Method of Handling Curvilinear Correlation for any Number of Variables, Journal of American Statistical Association, Vol. XIX, New Series No. 148, pp. 431–53, December 1924.

To facilitate the work, these values, +6.1, −1.8, +1.1, +23.0, and 28.4, may be recorded systematically, as in table 2, columns 5 to 9.

TABLE 2.—DETERMINATION OF RESIDUALS FROM THE LINEAR REGRESSION

CORN ON NORTH CAROLINA FARMS

Year	Independent variables			Calculation of residuals							
	Price of		Stocks of corn	Values of			Constant	Sum of values $+a$	Acres corn	Residuals	
	Corn	Cotton								$X_1 - X_1'$	Squared
	X_2	X_3	X_4	$0.0842X_2$	$-0.2312X_3$	$0.0971X_4$	a	X_1'	X_1	z	z^2
1	72	8	11	6.1	−1.8	1.1	23.0	28.4	27	−1.4	1.96
2	45	9	16	3.8	−2.1	1.6	23.0	26.3	26	−0.3	0.09
3	50	14	17	4.2	−3.2	1.7	23.0	25.7	27	+1.3	1.69
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
23	47	20	29	4.0	−4.6	2.8	23.0	25.2	23	−2.2	4.84
24	61	14	18	5.1	−3.2	1.7	23.0	26.6	24	−2.6	6.76
25	42	11	19	3.5	−2.5	1.8	23.0	25.8	24	−1.8	3.24
Total	—	—	—	—	—	—	—	—	—	+0.5	51.01
Average	—	—	—	—	—	—	—	—	—	+0.02	2.040

$\sigma_1^2 = 3.60.$

$$R_{1.234} = \sqrt{1 - \frac{S^2_{1.234}}{\sigma_1^2}} = \sqrt{1 - \frac{2.040}{3.6}} = \sqrt{1 - 0.5667} = \sqrt{0.4333} = 0.658.$$

Rather than calculate the complete equation for each year at one time, the simplest procedure is to multiply the first regression coefficient, $b_{12.34} = 0.0842$, by each of the values of X_2 in the second column of table 2. The products are recorded in table 2, column 5, under the heading "Values of $0.0842X_2$." The values of X_3 are then multiplied by the second regression coefficient, $b_{13.24} = -0.2312$, and recorded in column 6; and the values of X_4, by $b_{14.23} = 0.0971$, and recorded in column 7. The sums of these three products plus the constant 23.0 give the estimated prices, X_1'.

The residual[8] for the first year was $X_1 - X_1' = 27.0 - 28.4 = -1.4$;

---

[8] The sum of all the residuals, +0.5, was useful in checking the calculations. The average residual, 0.5 + 25 = 0.02, was small enough to be explained by the rounding of decimals. If the average residual were large, an error in calculation would be indicated.

The residuals were squared and entered in the last column of table 2.

and for the second year, $26.0 - 26.3 = -0.3$ (entered in the next to the last column of table 2). The 25 residuals, z, are the primary purpose of the work in table 2. The residuals are squared only for use in computing $R_{1.234}$.

X_1X_2 Relationship. The third step consists of plotting on a graph the relationship between each independent variable and the dependent variable. For example, the linear relationship between X_1 and X_2 was plotted as a broken line (figure 2). To determine the equation of this line from the net regression,

$$X_1 = 0.0842X_2 - 0.2312X_3 + 0.0971X_4 + 23.01$$

the variables X_3 and X_4 were held constant at their averages, $AX_3 = 11.92$ and $AX_4 = 22.96$.

$$\begin{aligned} X_1 &= 0.0842X_2 - 0.2312(11.92) + 0.0971(22.96) + 23.01 \\ X_1 &= 0.0842X_2 - 2.76 + 2.23 + 23.01 \\ X_1 &= 0.0842X_2 + 22.48 \end{aligned}$$

To plot this straight line, the estimated acreage, X_1', was calculated for two arbitrary values of X_2, 30 and 70.

$$\text{When } X_2 = 30,\ X_1' = 25.0. \quad \text{When } X_2 = 70,\ X_1' = 28.4$$

These two points were plotted as small circles, o, and the broken line connecting them was drawn (figure 2). At this stage, figure 2 contains only a broken line.

The fourth step consists of plotting the residuals (next to the last column in table 2) about the broken line in figure 2. For example, the value of X_2 for the first year was 72, and the residual was $z = -1.4$. The first year was plotted horizontally according to X_2, and vertically above or below the broken line according to the size and sign of z. The value of X_2 for the first year, 72, was located far to the right on the horizontal scale, X_2. The plotted point for the first year, designated by "1," is directly above 72 on the horizontal scale, X_2. This point is -1.4

Since $\Sigma z^2 = 51.01$, the standard error of estimate was

$$S_{1.234}^2 = \frac{\Sigma z^2}{N} = \frac{51.01}{25} = 2.040$$

and the multiple correlation coefficient,

$$R_{1.234} = \sqrt{1 - \frac{S_{1.234}^2}{\sigma_1^2}} = \sqrt{1 - \frac{2.040}{3.6}} = \sqrt{1 - 0.5667} = \sqrt{0.4333} = 0.658$$

The only purpose of calculating this R is to give another check on the calculation of residuals, 0.658 (above), as compared with 0.666 (table 1).

from the line of regression, that is, 1.4 below the broken line. The size of the scale for this residual, which is a deviation in X_1, is the same as the vertical scale of X_1 to the left of the chart (figure 2). However, the positions of the two vertical scales have no relationship. The position of the vertical scale to the left of figure 2 is *fixed*, whereas the vertical scale for the residual is constantly *moving up* or *down*, depending on the level of the broken line. The fixed scale refers only to the broken line and the curve to be drawn later.

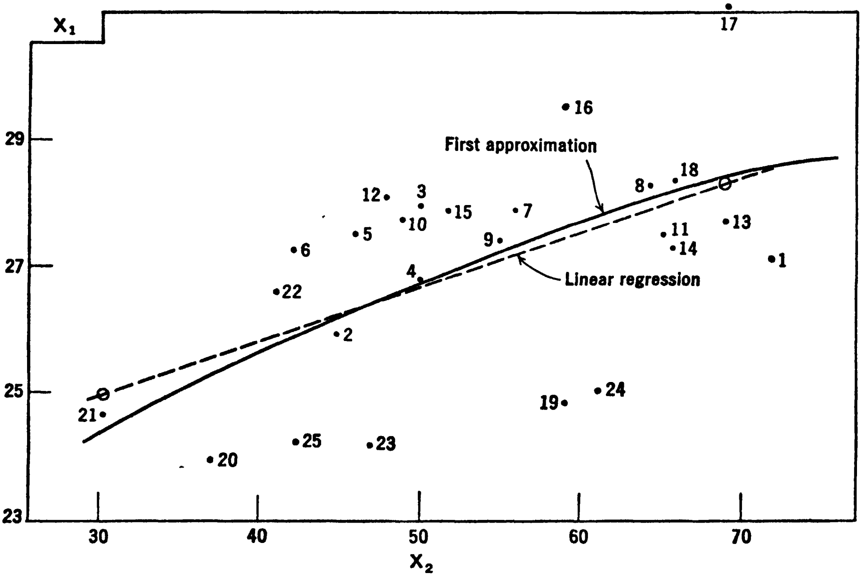


FIGURE 2.—FIRST APPROXIMATION CURVE FOR THE X_1 AND X_2 RELATIONSHIP

ACRES OF CORN, X_1, AND PRICE OF CORN THE PRECEDING YEAR, X_2

Approximation from Linear Regression

The broken line was the best-fitting straight line describing the net relationship between price and acreage of corn. The 25 scattered points were the residuals, plotted about the broken straight line. Based on the scatter of these points about the straight line, the solid curve was approximated for the purpose of showing the relationship more accurately.

The residual, z, is always plotted above or below the broken line with no reference to the fixed X_1 scale at the left.[9] In order to eliminate this apparent confusion, a separate scale for the residuals should be made on a piece of graph paper. The second scale in this case should

---

[9] In terms of the fixed X_1 scale at the left, the plotted points describe the total variability in X_1 unexplained by the relationship

$$X_1 = a' + b_{13.24}X_3 + b_{14.23}X_4$$

range from about +3.0 to 0 to −3.0 and should be the same size as the fixed scale in X_1 from 23 to 29, six points. To plot the first year with the new scale, the zero point would be placed on the broken line where $X_2 = 72$; and the point at −1.4, marked "1." Since the value of X_2 for the second year was 45, the scale was slid down the broken line to the left until the point 45 on the horizontal scale, X_2, was reached. With the zero point resting on the broken line, the point at −0.3 was marked "2." For the third year, the scale was slid up the broken line to the right so that $X_2 = 50$. The residual being +1.3, the point was plotted above the line and marked "3." The remaining 22 years were plotted by the same method. At this stage, the chart contains a broken line and 25 points numbered 1 to 25.

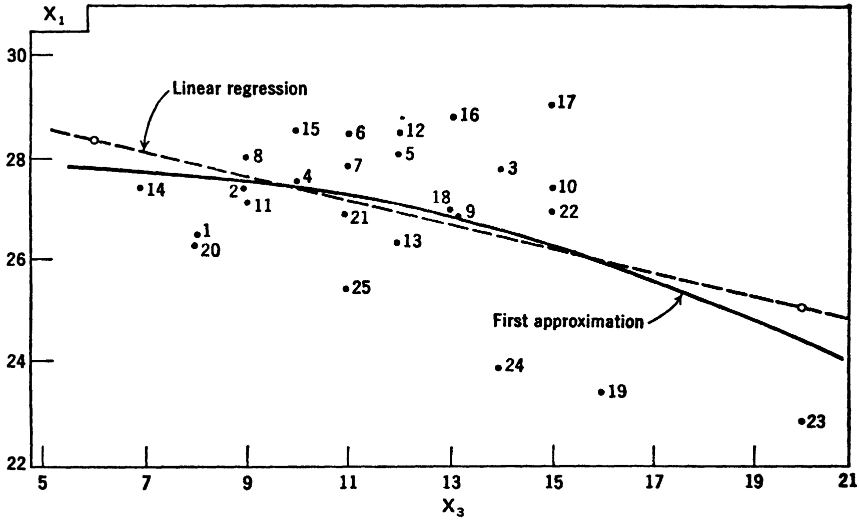


FIGURE 3.—FIRST APPROXIMATION CURVE FOR THE X_1 AND X_3 RELATIONSHIP

ACRES OF CORN, X_1, AND PRICE OF COTTON THE PRECEDING YEAR, X_3

Approximation from Linear Regression

The broken line describes the linear relationship between the price of cotton the preceding year and the acreage of corn planted. The solid curve presumably describes the relationship more accurately.

The first four steps are mechanical. In the fifth step, personal judgment appears for the first time and affects the results. The problem is to improve on the broken straight line in describing the relationship shown by the scatter of the 25 points. Obviously, no other *straight* line would describe the relationship so accurately as the broken line. Perhaps a *curve* of some type would describe the relationship more accurately.

After examination of the scatter in the 25 points, an *approximation* curve was drawn freehand (figure 2). Since the scatter among the points was considerable and the relationship was not decidedly curvilinear, it was thought safest to draw the first approximation not greatly different in position or shape from the broken straight line. The curve was also conservative in having only one bend.

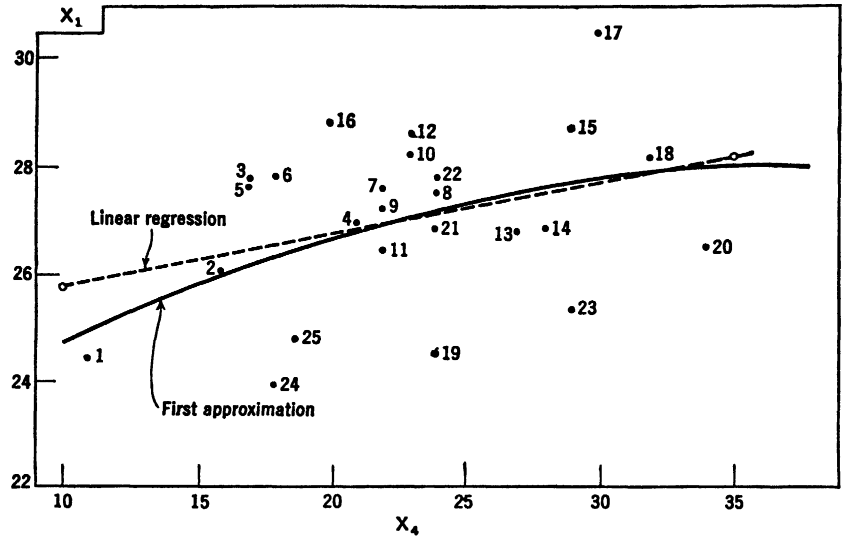


FIGURE 4.—FIRST APPROXIMATION CURVE FOR THE X_1 AND X_4 RELATIONSHIP

ACRES OF CORN, X_1, AND STOCKS OF CORN, X_4

Approximation from Linear Regression

The scatter diagram suggested that a curve increasing at a decreasing rate might describe the relationship more accurately than the straight broken line.

X_1X_3 and X_1X_4 Relationships. Next, the relationship between X_3 and X_1 was examined. In the multiple linear regression equation, X_2 and X_4 were held constant at their averages, and the net relationship between X_3 and X_1 was determined.

$$X_1 = 0.0842(53.6) - 0.2312X_3 + 0.0971(22.96) + 23.0$$
$$X_1 = 4.5 - 0.2312X_3 + 2.2 + 23.0$$
$$X_1 = 29.7 - 0.2312X_3$$

This straight line is the broken line plotted in figure 3. The residuals calculated in table 2 were then plotted above and below the broken line at the points corresponding to the values of X_3. For the first year, $z = -1.4$ and $X_3 = 8$. With the use of the deviation scale already constructed, the point 1 was placed at -1.4 below the broken line where $X_3 = 8$. Similarly, point 3 was placed at $+1.3$ above the broken line where $X_3 = 14$. The residuals for all the 25 years were plotted in this manner.

The next problem was to draw a curve which approximated the relationship between X_3 and X_1 more closely than the broken straight line. A first approximation curve, declining at an increasing rate, was drawn (figure 3).

The relationship between X_1 and X_4 was next examined (figure 4). The broken line was plotted from the net relationship

$$X_1 = 0.0971X_4 + 24.8$$

The residuals were plotted about the broken line at the positions corresponding to X_4 (table 2). The curved solid line represents the first curvilinear approximation to the relationship (figure 4).

Index of Multiple Correlation. The sixth step is to combine the three curves into one multiple relationship. The acres of corn, X_1, were estimated from each of the three curvilinear relationships by reading the values from the curves. The first year, the price of corn, X_2, was 72. The estimated acreage, $X_1' = 28.6$, was read from the solid curved line (figure 2). For the second year, X_2 was 45, and X_1' was 26.2. These estimated acreages, X_1', were tabulated in an orderly fashion in table 3, column 5, headed $f'(X_2)$.

Similarly, the estimated values of X_1 were determined from the X_3 relationship (figure 3). For the first year, the price of cotton, X_3, was 8 (table 3). The estimated acreage, $f'(X_3)$, 27.7, was read from the solid curve in figure 3. For the second year, X_3 was 9 and $f'(X_3) = 27.6$.

The estimated values of X_1 from the X_4 relationship shown in figure 4 were tabulated in table 3, column 7, headed $f'(X_4)$.

On the basis of the three curvilinear relationships, the estimated values of X_1 for the first year were 28.6, 27.7, and 24.9 (table 3). The sum of these estimates was 81.2, which was entered in column 8, headed $\Sigma f'$. Similarly, the estimates for each year were summed and entered in column 8. The sum for this column was 2,022.3. Its average, 80.9, was somewhat larger than the average of the actual, $AX_1 = 27$. The difference between the two averages, -53.9, may be regarded as a constant, a'. This constant was added to the sum of the three estimated values of X_1 for each year to obtain a new estimated value that would average 27, the same as the actual, AX_1. This constant,[10] $a' = -53.9$,

---

[10] This constant is a part of a curvilinear regression equation:

$$X_1' = f'(X_2) + f'(X_3) + f'(X_4) + a'$$

TABLE 3.—DETERMINATION OF THE RESIDUALS FROM THE MULTIPLE RELATIONSHIPS BASED ON THE FIRST APPROXIMATION CURVES

CORN ON NORTH CAROLINA FARMS

Year	Independent variables			Calculation of residuals								
				Estimated values of X_1				Constant	Estimated	Actual	Residuals	
											$X_1 - X_1'$	Squared
	X_2	X_3	X_4	$f'(X_2)$	$f'(X_3)$	$f'(X_4)$	$\Sigma f'$	a'	X'_1	X_1	z'	$(z')^2$
1	72	8	11	28.6	27.7	24.9	81.2	−53.9	27.3	27	−0.3	0.09
2	45	9	16	26.2	27.6	26.0	79.8	−53.9	25.9	26	+0.1	0.01
3	50	14	17	26.7	26.7	26.1	79.5	−53.9	25.6	27	+1.4	1.96
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
23	47	20	29	26.4	24.5	27.7	78.6	−53.9	24.7	23	−1.7	2.89
24	61	14	18	27.8	26.7	26.3	80.8	−53.9	26.9	24	−2.9	8.41
25	42	11	19	25.8	27.3	26.4	79.5	−53.9	25.6	24	−1.6	2.56
Total	—	—	—	—	—	—	2,022.3	—	675.0	675	0	44.20
Average	—	—	—	—	—	—	80.9	—	27.0	27	0	1.768

Constant
$a' = AX_1 - A(\Sigma f') = 27 - 80.9 = -53.9$

$$\rho_{1.234} = \sqrt{1 - \frac{S^2_{1.234}}{\sigma^2_1}} = \sqrt{1 - \frac{1.768}{3.600}} = \sqrt{1 - 0.4911} = \sqrt{0.5089} = 0.713$$

was recorded for each year in column 9. It was added to $\Sigma f'$ for each year, and these sums were recorded in column 10, headed "Estimated X'_1."

Thus far, the sixth step has been merely an orderly tabulation for the determination of the estimated values, X'_1, from the curvilinear regression equation expressed by $X'_1 = f'(X_2) + f'(X_3) + f'(X_4) + a'$. For the first year, this equation read

$$\begin{aligned} X'_1 &= 28.6 + 27.7 + 24.9 - 53.9 \\ &= 27.3 \end{aligned}$$

The plotting of data, drawing curves, and tabulating estimated values of X_1 discussed in detail should not obscure one of the important parts of the analysis, namely, the determination of ρ and ρ^2. The residuals, $X_1 - X'_1$, were calculated, squared, and entered in the last two columns of table 3. The average of the squared residuals, 1.768, was the

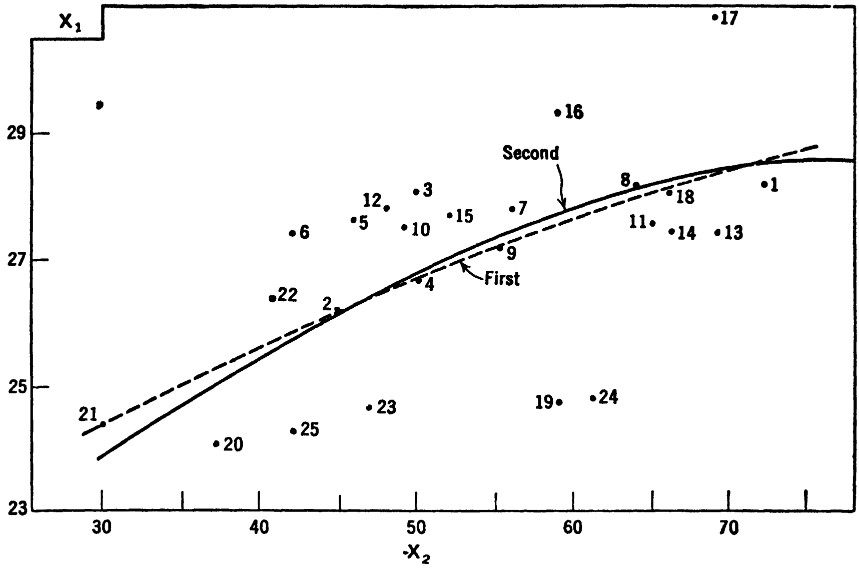


FIGURE 5.—SECOND APPROXIMATION CURVE FOR THE X_1X_2 RELATIONSHIP

ACRES OF CORN, X_1, AND PRICE OF CORN THE PRECEDING YEAR, X_2

Approximation from Linear Regression

The broken line was the first approximation curve transferred from figure 2. The 25 points were the residuals from the first approximation curves, table 3, plotted above and below the broken line. The solid line, the second approximation curve, was an attempt to describe the X_1X_2 relationship more accurately.

squared standard error of estimate. The index of multiple correlation was

$$\rho_{1.234} = \sqrt{1 - \frac{1.768}{3.600}} = \sqrt{0.5089} = 0.713$$

The work to this point may be summarized as follows:

For the linear relationship,

$$S^2_{1.234} = 2.040; \quad R_{1.234} = 0.658; \quad \text{and } R^2_{1.234} = 0.433$$

For the curvilinear relationship,

$$S^2_{1.234} = 1.768; \quad \rho_{1.234} = 0.713; \quad \text{and } \rho^2_{1.234} = 0.509$$

Apparently, the three curves described the multiple relationship more accurately than the straight lines. The index of correlation, $\rho_{1.234} = 0.713$, was 8 per cent greater than the multiple correlation coefficient, $R_{1.234} = 0.658$. From the curvilinear relation, it would appear that these prices of corn and cotton and stocks of corn account for 51 per cent of the variation in acres of corn. This per cent determination, 51, was greater than that from linear analysis, 43.

The squared standard error of estimate which is a measure of the amount of scatter or unexplained variability in the price of corn, X_1, was less for the curvilinear than for the linear relationship.

The curvilinear analysis to this point is complete in itself, provided that the results are satisfactory.

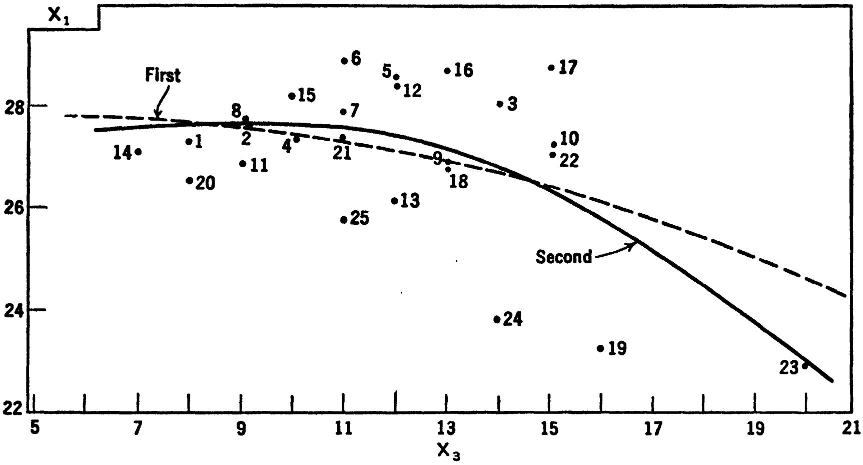


FIGURE 6.—SECOND APPROXIMATION CURVE FOR THE X_1X_3 RELATIONSHIP

ACRES OF CORN, X_1, AND PRICE OF COTTON THE PRECEDING YEAR, X_3

Approximation from Linear Regression

The second approximation curve was believed to describe the relation between X_1 and X_3 more accurately than the first.

Second Approximations

The student who is not satisfied with the results of the first approximations may wish to increase the accuracy of the curves and the index of correlation, ρ. Since $\rho_{1.234} = 0.713$ was only a little larger than $R_{1.234} = 0.658$, the authors felt that the first approximation curves could be improved.

The procedure for a set of second approximation curves was almost a repetition of the first approximation.

X_1X_2 *Relationship.* The first step was to reproduce the first approximation curve for the X_1X_2 relationship. The solid line in figure 2 was traced as a broken line in figure 5.

The second step was to plot the residuals from the first approximation, z', about this broken line. For the first year, $X_2 = 72$ and $z' = -0.3$

(table 3, second and next to last columns). Using the deviation scale, the point 1 for the first year was placed at −0.3, which was 0.3 below the broken line at the point corresponding to $X_2 = 72$ on the horizontal axis. The other 24 residuals, z', in table 3 were similarly plotted about the broken curve (figure 5).

The third step was to draw another curve which fitted the scatter better than the broken, first approximation curve. In the example, the second approximation was little different from the first.

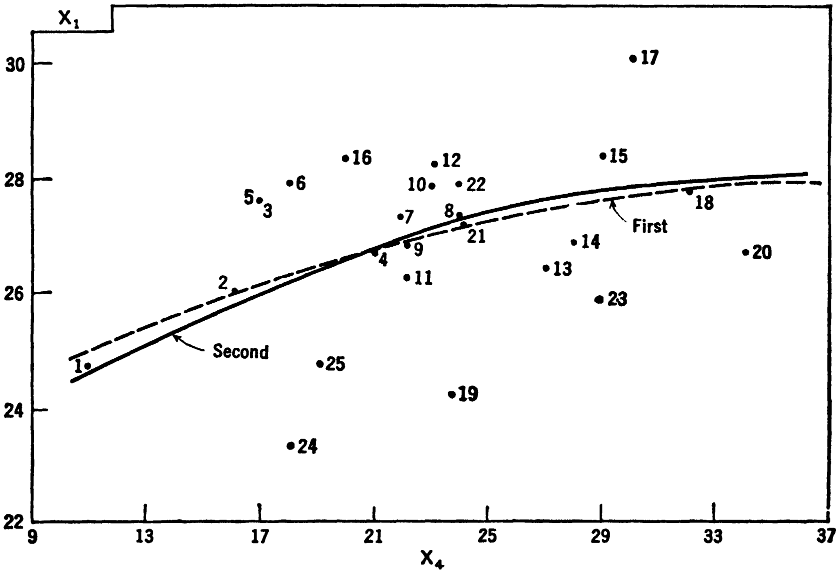


FIGURE 7.—SECOND APPROXIMATION CURVE FOR THE X_1X_4 RELATIONSHIP

ACRES OF CORN, X_1, AND STOCKS OF CORN, X_4

Approximation from Linear Regression

The second approximation curve was slightly more curvilinear than the first.

X_1X_3 and X_1X_4 Relationships. The same procedure was followed for the X_1X_3 relationship. The solid line in figure 3 was transferred to figure 6 as a broken line. The residuals, z', were plotted about this broken curve. A second approximation curve was drawn.

The X_1X_4 relations were analyzed in the same way, and a second approximation curve was drawn (figure 7).

Index of Multiple Correlation. The next step was to combine the three new curves and determine ρ. The three sets of estimated values of X_1 were read from the three second approximation curves.

For the X_1X_2 relationship, the procedure was as follows: For the first year, X_2 was 72, and the estimated acreage of corn, 28.6, was the value of the curve where the price of corn was 72 (figure 5). The value 28.6 was recorded in table 4, column 5, headed $f''(X_2)$.

From the X_1X_3 relationship, the estimated value of X_1 was 27.6 for the first year when $X_3 = 8$ (figure 6). This value was recorded in table 4, column 6, headed $f''(X_3)$.

TABLE 4.—DETERMINATION OF THE RESIDUALS FROM THE SECOND APPROXIMATION CURVES

CORN ON NORTH CAROLINA FARMS

Year	Independent variables			Calculation of residuals								
				Estimated values of X_1				Constant	Estimated	Actual	Residuals	
											$X_1 - X_1''$	Squared
	X_2	X_3	X_4	$f''(X_2)$	$f''(X_3)$	$f''(X_4)$	$\Sigma f''$	a''	X''_1	X_1	z''	$(z'')^2$
1	72	8	11	28.6	27.6	24.6	80.8	−54.1	26.7	27	+0.3	0.09
2	45	9	16	26.1	27.7	25.8	79.6	−54.1	25.5	26	+0.5	0.25
3	50	14	17	26.7	26.9	26.0	79.6	−54.1	25.5	27	+1.5	2.25
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
23	47	20	29	26.4	23.0	27.8	77.2	−54.1	23.1	23	−0.1	0.01
24	61	14	18	27.9	26.9	26.3	81.1	−54.1	27.0	24	−3.0	9.00
25	42	11	19	25.6	27.6	26.5	79.7	−54.1	25.6	24	−1.6	2.56
Total	—	—	—	—	—	—	2,026.9	—	674.4	675	0	39.90
Average	—	—	—	—	—	—	81.1	—	27.0	27	0	1.5960

Constant
$a'' = AX_1 - A(\Sigma f'') = 27.0 - 81.1 = -54.1$

$$\rho_{1.234} = \sqrt{1 - \frac{S^2_{1\ 234}}{\sigma^2}} = \sqrt{1 - \frac{1.5960}{3.600}} = \sqrt{1 - 0.4433} = \sqrt{0.5567} = 0.746$$

The X_1X_4 relationship was treated in the same manner.

For each year, the three estimates for X_1 were summed, and the totals entered in column 8, headed $\Sigma f''$. For the first year, the estimates were 28.6, 27.6, 24.6; and their sum, 80.8. The average for this column 8, $\Sigma f''/N = 81.1$, was subtracted from the average of X_1, 27.0, to obtain a constant, $a'' = -54.1$. For each year, this constant was added to $\Sigma f''$ to determine the estimated value of X_1 based on the relationship $X_1'' = f''X_2 + f''X_3 + f''X_4 + a''$. For the first year,

$$X_1'' = 28.6 + 27.6 + 24.6 - 54.1$$
$$X_1'' = 80.8 - 54.1 = 26.7$$

The residuals, the differences between the actual and estimated values of X_1, were calculated, squared, and entered in the last column of table 4. The index of correlation from the second approximation curves was $\rho_{1.234} = 0.746$, somewhat higher than that from the first approximation curves ($\rho_{1.234} = 0.713$). Since rho is merely an index of the closeness with which the curves fit, the second approximation curves were presumably more accurate than the first.

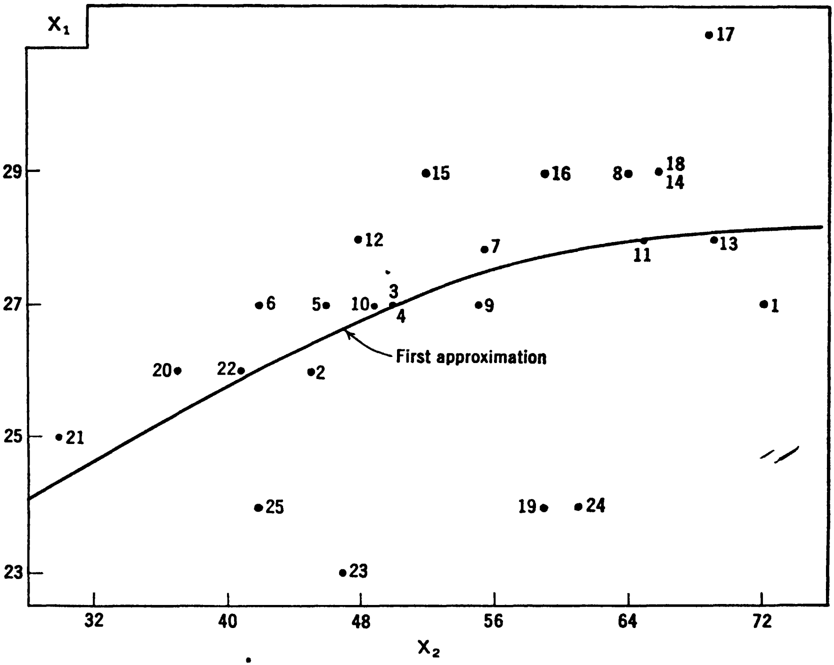


FIGURE 8.—FIRST SHORT-CUT APPROXIMATION CURVE FOR X_1X_2 RELATIONSHIP

ACRES OF CORN, X_1, AND PRICE OF CORN THE PRECEDING YEAR, X_2

Short-Cut Method of Approximation

The numbered points were the paired acres and prices of corn for 25 different years. The solid line was an attempt to describe the average change in the acreage of corn, X_1, with changes in the prices of corn.

If the second approximation curves are not satisfactory, the student might repeat the processes and obtain third or even fourth approximation curves. If the work is carefully done, the accuracy of the curves and the size of ρ will not be increased greatly after the second approximation.

SHORT-CUT METHOD OF APPROXIMATION

Curvilinear approximations from linear regressions usually give quite satisfactory results, but a considerable amount of work is involved. Bean[11] developed a method of approximating curvilinear relationships without the previous determination of linear regressions.

Relation of Dependent to Most Important Independent Variable

The first step in this method is to plot the observations on a graph where the vertical scale is the dependent variable, X_1; and the horizontal scale, one of the independent variables, X_2, X_3, or X_4. It is generally advisable to consider first the independent variable which is most closely associated with the dependent variable, X_1. It was assumed that X_1, the acres of corn, was more closely associated with X_2, the price of corn, than with X_3 or X_4, the price of cotton and stocks of corn.

The X_1X_2 relationship was plotted in figure 8 from the original values of X_1 and X_2 indicated in table 1, page 214. The resulting scatter indicated the relationship between X_1 and X_2, not considering X_3 or X_4 in any way.

A curve of relationship was drawn through the scatter and labeled "first approximation" (figure 8). In general, the curve should be drawn so that the squares of the residuals will be as small as possible. The student is less likely to make mistakes if he adheres to simple curves with only one bend.

Relation to Other Independent Variables

The second step was to plot the residuals[12] from the curve in figure 8 with X_3, which was considered the next most important independent variable (figure 9). The horizontal scale of figure 9 was in terms of actual values of X_3; and its vertical scale, in terms of residuals from

---

[11] Bean, L. H., A Simplified Method of Graphic Curvilinear Correlation, Journal of the American Statistical Association, Volume XXIV, New Series, No. 168, pp. 386–397, December 1929.

[12] Ordinarily, the values of the residuals are read from the curve (figure 8) and plotted with X_3 directly on the next chart (figure 9). Some students may find it advisable to list the values of X_3 from table 1 and then tabulate the corresponding residuals for the X_1X_2 relationship read from figure 8, as follows:

YEAR	X_3	z	YEAR	X_3	z
1	8	−1.2	23	20	−3.6
2	9	−0.4	24	14	−3.9
3	14	0	25	11	−2.0

The above residuals, z, were the differences between the actual acreage of corn and that estimated from the price of corn. They may be expressed algebraically as follows:

$$z = X_1 - f(X_2).$$

the curve, figure 8. For the first year, the price of cotton, X_3, was 8, and the residual read from figure 8 was −1.2. Point 1 was placed far to the left below the broken zero line (figure 9). The broken line has no significance except to aid in plotting. For the second year,[12] the values of X_3 and z were 9 and −0.4, respectively. The point 2 appears to the left below the broken line in figure 9.

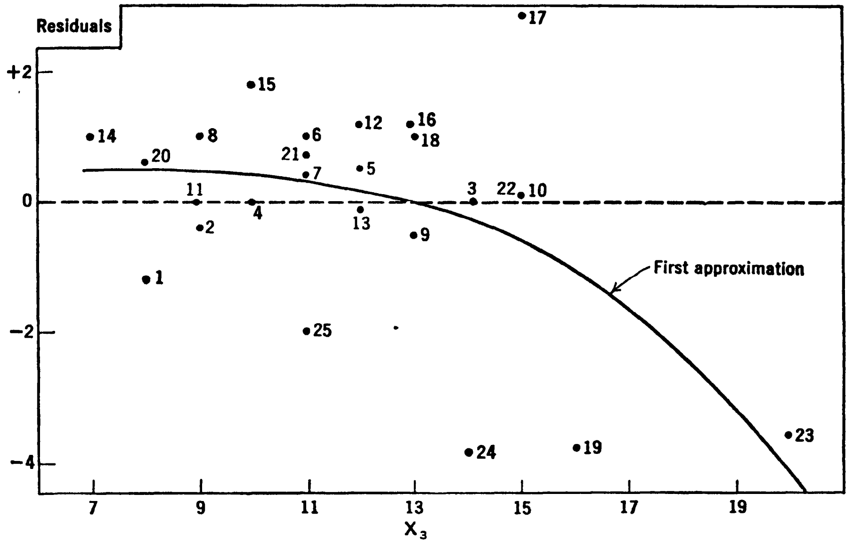


FIGURE 9.—RESIDUALS[13] FROM THE X_1X_2 APPROXIMATION CURVE PLOTTED WITH X_3

DIFFERENCES BETWEEN THE ACTUAL ACREAGE OF CORN AND THE ACREAGE ESTIMATED FROM THE PRICE OF CORN PLOTTED WITH THE PRICE OF COTTON, X_3

Short-Cut Method of Approximation

The 25 numbered points show the relationship between the price of cotton, X_3, and the acres of corn, X_1, not explained by the price of corn, X_2. The broken zero line has no value except to assist in plotting points. The solid line was an attempt to describe the unaccounted-for variation in X_1 in terms of X_3.

While following the detailed description of the process, the student should keep in mind that z is the amount of variability in the acres of corn, X_1, that was not accounted for by the price of corn, X_2. These residuals were plotted with X_3 to discover whether any of the variability not accounted for by X_2 could be ascribed to X_3. The relationship between X_3 and these residuals, z, was drawn as a solid curve that decreased at an increasing rate (figure 9). As the price of cotton, X_3, rose, the acreage of corn, X_1, declined at an increasing rate.

The third step was to plot the residuals from the curve in figure 9

---

[13] Residuals in X_1 from figure 8.

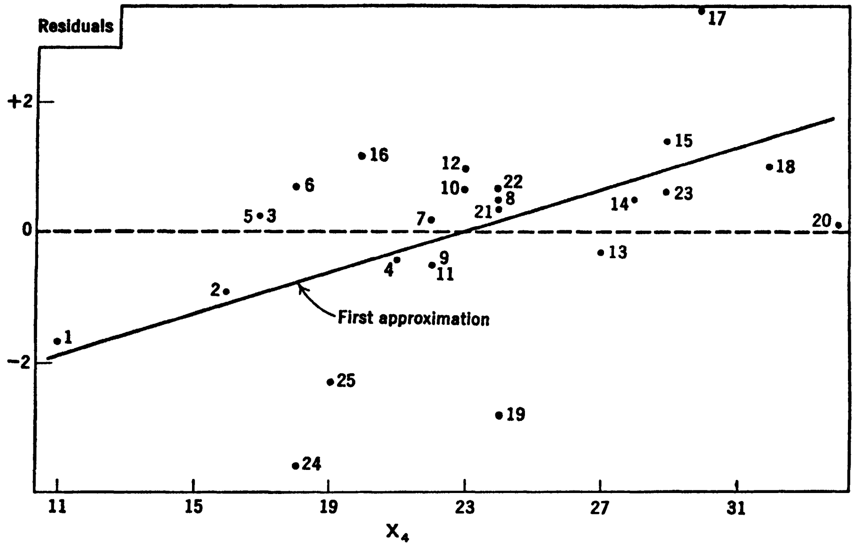


FIGURE 10.—RESIDUALS[14] FROM THE X_1X_3 APPROXIMATION CURVE PLOTTED WITH X_4

DIFFERENCES BETWEEN THE ACTUAL ACREAGE OF CORN AND THE ACREAGE ESTIMATED FROM THE PRICES OF CORN AND COTTON, z, PLOTTED WITH THE STOCKS OF CORN, X_4

Short-Cut Method of Approximation

The 25 numbered points show the relationship between the stocks of corn, X_4, and the variability in the acres of corn, X_1, unexplained by the prices of corn and cotton, X_2 and X_3. The solid line was an attempt to describe this relationship.

with X_4, the last independent variable (figure 10). The horizontal scale was in terms of the actual values of X_4. The vertical scale was in terms of residuals[15] from figure 9. For the first year,[15] the value of X_4 was 11, and z was -1.7. The point numbered 1 appears far to the left, below the broken zero line, figure 10. For the second year, the values of X_4 and z were 16 and -0.9, and the point appears below the broken zero line, to the left of figure 10. A straight line was drawn to represent the

---

[14] Residuals from figure 9.

[15] The values of X_4 and the residuals, from table 1 and figure 9, respectively, were as follows:

YEAR	X_4	z	YEAR	X_4	z
1	11	−1.7	23	29	+0.6
2	16	−0.9	24	18	−3.6
3	17	+0.3	25	19	−2.3

These residuals, which were read from figure 9, were plotted with X_4 in figure 10. These residuals, z, may be expressed algebraically as follows:

$$z = X_1 - f(X_2) - f(X_3, \text{ after considering } X_2)$$

average relationship indicated by the 25 points. This line describes the relationship between X_4 and the variability in X_1 not explained by X_2 and X_3. It was not, as is sometimes assumed, a description of the net relationship between X_1 and X_4. The effects of X_2 and X_3 on X_1 have been considered, but the effects of X_4 on the X_1X_2 and X_1X_3 relationships have not been considered.

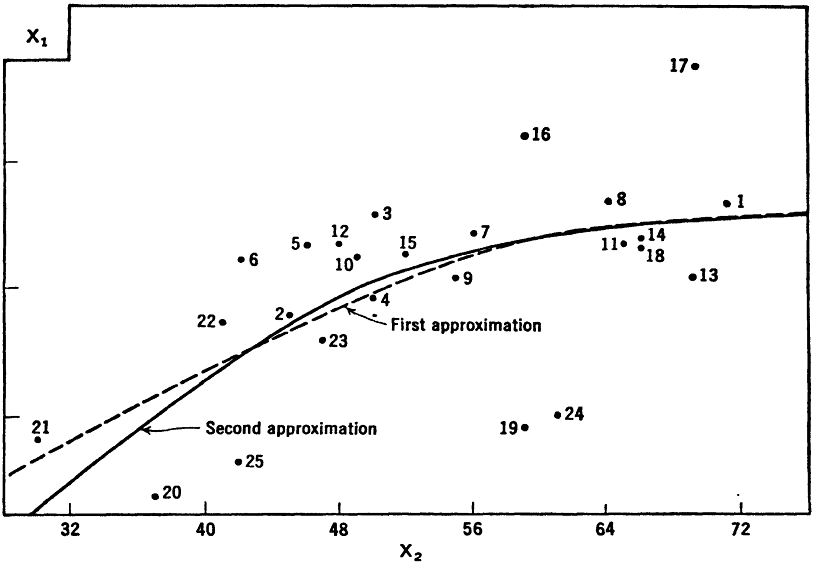


FIGURE 11.—RESIDUALS[16] FROM X_1X_4 APPROXIMATION PLOTTED ABOUT THE X_1X_2 APPROXIMATION CURVE,[17] WITH RESPECT TO X_2

Short-Cut Method of Approximation

The first approximation curve considered only the total relation between X_1 and X_2 ignoring all other factors. The second approximation curve for X_1 in terms of X_2 considered the interrelationship between X_2, and X_3 and X_4.

Elimination of Interrelationships

The fourth step was to consider the effects of X_3 and X_4 on the X_1X_2 relationship. This is done by plotting the residuals[18] from figure 10

---

[16] Residuals from figure 10. [17] Traced from figure 8.

[18] The residuals measured the variability in X_1 unexplained by X_2, X_3, and X_4 considered in that order. The values of X_2 and the residuals, from table 1 and figure 10, respectively, were as follows:

YEAR	X_2	z	YEAR	X_2	z
1	72	+0.2	23	47	−0.4
2	45	+0.2	24	61	−2.8
3	50	+1.2	25	42	−1.7

These residuals which were read from figure 10 were plotted with X_2 in figure 11. These residuals may be expressed algebraically as follows:

$$z = X_1 - f(X_2) - f(X_3\text{, after considering } X_2) - f(X_4\text{, after } X_3\text{, after } X_2)$$

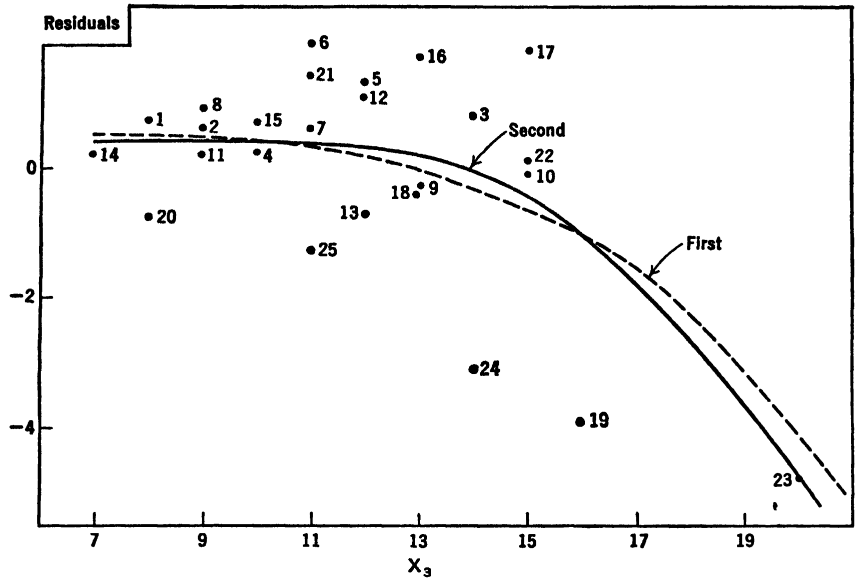


FIGURE 12.—RESIDUALS[19] FROM THE X_1X_2 SECOND APPROXIMATION PLOTTED ABOUT THE X_1X_3 FIRST APPROXIMATION CURVE,[20] WITH RESPECT TO X_3

Short-Cut Method of Approximation

The first X_1X_3 approximation curve considered only the relation between X_1 and X_3, eliminating the effect of X_2, but ignoring the effects of X_4 on X_3 and of X_3 and X_4 on X_2. The second approximation curve considered these interrelationships.

with respect to X_2 about the X_1X_2 first approximation curve shown in figure 8. The X_1X_2 curve in figure 8 was traced as a broken curve on figure 11. The residuals from figure 10 were then plotted about the broken curve.[21] For the first year, $X_2 = 72$, and the residual, $z = +0.2$, was plotted above the broken curve where $X_2 = 72$. The point was numbered 1 and lies far to the right in figure 11. After the 25 points are plotted, it should be observed whether a second approximation curve might not fit the scatter better than the first. The solid line is such a second approximation. The two approximations were different because in the first only the gross or total relation between X_1 and X_2 was considered, while in the second the effects of X_3 and X_4 on X_2 and X_1 were also considered (figure 11). The second approximation curve in figure 11 was probably a close approximation to the *net* relationship between X_1 and X_2.

---

[19] Residuals from figure 11.

[20] Traced from figure 9.

[21] In figures 9 and 10, the residuals were plotted about the broken zero line; in figure 11 they were plotted about the curve transferred from figure 8.

The fifth step was to consider the effects of X_4 and the revision of the X_1X_2 curve on the relationship between X_1 and X_3. The first approximation curve for the X_1X_3 relationship was traced from figure 9 as a broken curve on figure 12.

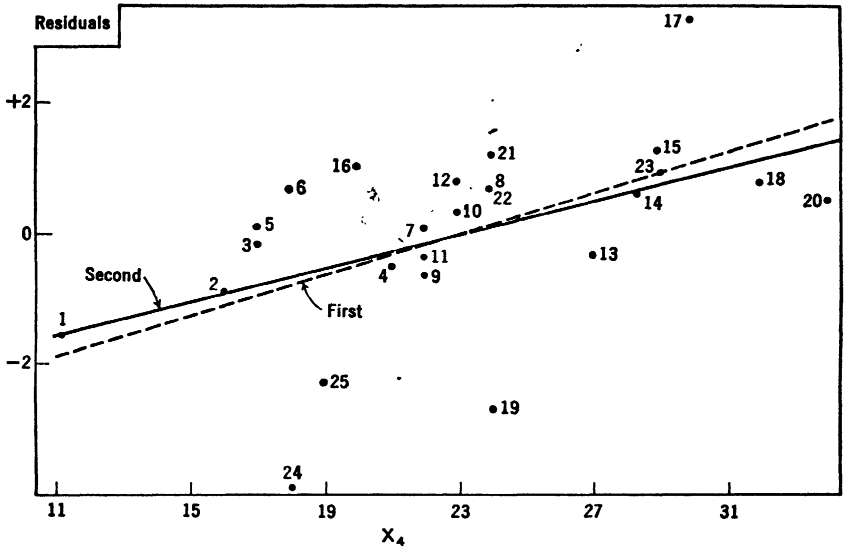


FIGURE 13.—RESIDUALS[22] FROM THE X_1X_3 SECOND APPROXIMATION PLOTTED ABOUT THE X_1X_4 FIRST APPROXIMATION,[23] WITH RESPECT TO X_4

Short-Cut Method of Approximation

Unlike the first approximation, the second approximation curve considered the effects of X_4 on the X_1X_2 and X_1X_3 curves.

The residuals[24] from the second approximation curve in figure 11 were plotted about the broken curve in figure 12 with respect to X_3. For the

[22] Residuals from figure 12. [23] Traced from figure 10.

[24] The values of X_3 and the residuals, from table 1 and figure 11, respectively, were as follows:

YEAR	X_3	z	YEAR	X_3	z
1	8	+0.2	23	20	−0.6
2	9	+0.1	24	14	−2.8
3	14	+1.1	25	11	−1.6

These residuals were read from figure 11 and plotted about the first approximation curve with respect to X_3 in figure 12. The residuals may be expressed algebraically as follows:

$$z = X_1 - f(X_3, \text{ after } X_2) - f(X_4, \text{ after } X_3, \text{ after } X_2) - f(X_2, \text{ after } X_4, \text{ after } X_3, \text{ after } X_2)$$

first year, $X_3 = 8$ and $z = +0.2$. The point 1 was plotted 0.2 above the broken line where $X_3 = 8$ in figure 12. The other 24 points were plotted about the broken curve. The scatter was then examined to detect whether the factors previously not considered, X_4 and the X_1X_2 net relationship, changed the X_1X_3 relationship. The solid curve is the second approximation. It is somewhat different from the first because it takes into account the effects of X_4 and the X_1X_2 revision on the X_1X_3 relationship. Stated another way, the solid curve in figure 12 showed the average net effect of the price of cotton, X_3, on the acreage of corn, X_1, taking into consideration the effects of price of corn, X_2, and stocks of corn, X_4.

The sixth step was the reconsideration of the X_1X_4 relationship in the light of the revised X_1X_2 and X_1X_3 curves. The straight line in figure 10 was traced as the broken line in figure 13. The residuals[25] from the second approximation curve in figure 12 were plotted about this broken line in figure 13, with respect to X_4. The scatter was then examined, and the second approximation curve was drawn as a solid line.

Index of Correlation

The three second approximation curves may be regarded as the net relationships between the acres of corn and each of the other factors. The index of correlation from the multiple relationship was next in order. The residuals about the curves were carried forward from one graph to the next in the process of drawing new approximations. The residuals about any one new approximation represented the variability in X_1 not accounted for by all factors and interrelationships considered to that point. The residuals from the multiple relationship are those measured about the last approximation drawn, the second approximation in figure 13. The index of correlation, $\rho = 0.768$, was based on the squares of these residuals (table 5).

[25] The values of X_4 and the residuals, from table 1 and figure 12, respectively, were as follows:

YEAR	X_4	z	YEAR	X_4	z
1	11	+0.3	23	29	0
2	16	+0.2	24	18	−3.1
3	17	+0.8	25	19	−1.7

These residuals were read from figure 12 and plotted about the first approximation curve with respect to X_4 in figure 13. The residuals may be expressed algebraically as follows:

$$z = X_1 - f(X_4, \text{after } X_3, \text{after } X_2) - f(X_2, \text{after } X_4, \text{after } X_3, \text{after } X_2) - f(X_3, \text{after } X_2, \text{after } X_4, \text{after } X_3, \text{after } X_2).$$

TABLE 5.—INDEX OF CORRELATION FROM RESIDUALS* ABOUT THE LAST APPROXIMATION CURVE, FIGURE 13

Year	z	z^2	Year	z	z^2	Year	z	z^2
1	0	0	10	+0.3	0.09	19	−2.9	8.41
2	0	0	11	−0.2	0.04	20	−0.9	0.81
3	+0.6	0.36	12	+0.8	0.64	21	+1.0	1.00
4	−0.3	0.09	13	−0.9	0.81	22	+0.5	0.25
5	+0.8	0.64	14	−0.1	0.01	23	+0.2	0.04
6	+1.3	1.69	15	+0.5	0.25	24	−3.2	10.24
7	+0.2	0.04	16	+1.4	1.96	25	−1.8	3.24
8	+0.5	0.25	17	+2.4	5.76			
9	−0.5	0.25	18	−0.3	0.09	Total		36.96
						Average		1.4784

$$\rho = \sqrt{1 - \frac{1.478}{3.600}} = \sqrt{0.5894} = 0.768$$

* These residuals may be expressed algebraically as follows:

$z = X_1 - f(X_2,$ after X_4, after X_3, after $X_2) - f(X_3$, after X_2, after X_4, after X_3, after $X_2) - f(X_4$, after X_3, after X_2, after X_4, after X_3, after $X_2)$

The residuals may also be obtained by adding together the three functional or estimated values of X_1. These values may be read from the three second approximation curves (figures 11, 12, and 13). For the first year, $X_1 = 27$ and

$f(X_2$, after X_4, after X_3, after $X_2)$ (figure 11) $= +28.2$
$f(X_3$, after X_2, after X_4, after X_3, after $X_2)$ (figure 12) $= +\ 0.4$
$f(X_4$, after X_3, after X_2, after X_4, after X_3, after $X_2)$ (figure 13) $= -\ 1.6$

$$z = AX_1 + f(X_2, \text{etc.}) + f(X_3, \text{etc.}) + f(X_4, \text{etc.})$$
$$z = 27 - 28.2 - 0.4 + 1.6$$
$$= 27 - 27 = 0$$

If the second approximation curves are not satisfactory, the process may be continued and third approximation curves obtained.

Significance of Each Approximation

The difficulties of following the calculations and understanding the principles involved are not peculiar to the short-cut method. Any problem that deals with four variables has many interrelationships of varying degrees of importance. Each step in the short-cut method has a definite purpose. An attempt will be made to summarize each of the six steps (table 6). Since the residuals were always known, the index of correlation and the coefficient of determination could be calculated after each step. This would enable the student to observe the additional variability explained by each step.

After the first step, it was apparent that not much of the squared variability in X_1 was explained by X_2, $\rho^2_{1.2} = 0.241$. The effects of X_3 and X_4 or of any interrelationships were not considered.

TABLE 6.—SUMMARY OF SHORT-CUT APPROXIMATION METHOD, STEP BY STEP

Step	Approximation, figure number and page	Relationship between	Independent variables considered	Interrelationships considered	As^2	ρ	ρ^2
1	8, p. 229	X_1, X_2	X_2	None	2.73	$\rho_{1.2} = 0.491$	0.241
2	9, p. 231	X_1, X_3	X_3, after eliminating effect X_2	X_2X_3, partly	2.04	$\rho_{1.23} = 0.659$	0.435
3	10, p. 232	X_1, X_4	X_4, after X_3, after X_2	X_2X_3, X_2X_4, and X_3X_4, partly	1.57	$\rho_{1.234} = 0.750$	0.563
4	11, p. 233	X_1, X_2	X_2, after X_4, after X_3, after X_2	X_2X_3 and X_2X_4, all; and X_3X_4, partly	1.56	$\rho_{1.234(2)} = 0.753$	0.568
5	12, p. 234	X_1, X_3	X_3, after X_2, after X_4, after X_3, after X_2	X_2X_3, X_3X_4, and X_2X_4, all	1.50	$\rho_{1.234(23)} = 0.763$	0.583
6	13, p. 235	X_1, X_4	X_4, after X_3, after X_2, after X_4, after X_3, after X_2	X_2X_4, X_3X_4, and X_2X_3, all	1.48	$\rho_{1.234(234)} = 0.768$	0.589

After the second step, it was apparent that 43.5 per cent of the variability in X_1 was explained by X_2 and X_3 ($\rho^2_{1.23} = 0.435$). This indicated that X_3 explained 19.4 per cent of the variability in X_1 in addition to that explained by X_2 $(0.435 - 0.241 = 0.194)$. In this step, the interrelationship between X_2 and X_3 was partly considered.

After the third step, it was apparent that 56.3 per cent of the variability in X_1 was explained by X_2, X_3, and X_4 ($\rho^2_{1.234} = 0.563$). This indicated that X_4 explained 12.8 per cent of the variability in X_1 in addition to that explained by X_2 and X_3. With this step, the interrelationships between X_2 and X_3, X_2 and X_4, and X_3 and X_4 had been *partly* considered.

In steps 4, 5, and 6, each relationship was reconsidered in the light of the interrelationships among the independent variables. The effect of the interrelationships on the coefficient of determination was observed.

After the fourth step, it was apparent that the reconsideration of the X_1X_2 relationship did not explain much additional variability ($\rho^2_{1.234(2)} = 0.568$, compared with $\rho^2_{1.234} = 0.563$). With this step, presumably *all* the interrelationships between X_2 and X_3 and between

X_2 and X_4 had been considered; and between X_3 and X_4, *partly* considered.[26]

The fifth and sixth steps presumably considered *all* the interrelationships, but raised the coefficient of determination only slightly ($\rho^2_{1.234(234)} = 0.589$, compared with $\rho^2_{1.234(2)} = 0.568$).

Interrelationships affect the shapes of curves as well as the size of rho. The effects of these interrelationships can be observed by comparing corresponding first and second approximation curves. The difference between the first and second X_1X_2 curves, shown by the broken and solid lines respectively in figure 11, indicates that the interrelationships did not materially change the shape of the curve describing the X_1X_2 relationship. The same was true for the X_1X_3 and X_1X_4 curves (figures 12 and 13).

The small changes in rho and the shapes of the curves indicated that the interrelationships were slight. The gross correlation coefficients between the independent variables were calculated and found to be small.[27] "Horse sense" would lead to the conclusion that there would be slight interrelationship between the price of corn and cotton, X_2X_3, and between the price of cotton and stocks of corn, X_3X_4. It might be expected that there would be some interrelationship between the price of corn and stocks of corn, X_2X_4. Had the interrelationship been larger, there would have been more decided changes in the curves and possibly greater increases in rho.

Guide to Drawing Approximations

In the discussion of the short-cut method, Bean's guide for the location of the approximation curves was omitted. When there are marked interrelationships between the independent variables, the use of this guide will yield the correct net regression curves with fewer approximations than would otherwise be needed. This device attempts to hold constant the effects of the other factors not considered in the relationship to which the curve is being drawn.

In the beginning of the analysis of the corn-acreage problem, X_1 was plotted with X_2, and a curve was drawn without the aid of this device (figure 8). Before this curve was drawn, the net relation, X_1 to X_2

---

[26] The interrelationship between X_2 and X_3 was said to be *all* considered because its effect on both the X_1X_2 and X_1X_3 relationships had been taken into account. Similarly, the interrelationship between X_2 and X_4 had been *all* considered. However, the X_3X_4 interrelationship had been only *partly* considered because its effect on the X_1X_4 relationship had been taken into account, but its effect on the X_1X_3 relationship had not.

[27] $r_{23} = -0.08$; $r_{24} = 0.03$; and $r_{34} = 0.15$.

eliminating the effects of X_3 and X_4, could have been examined from pairs of X_1 and X_2 for which the values of X_3 and X_4 were approximately the same. For instance, for the years 8 and 11, the values of X_3 and X_4 were approximately the same; 9 and 9, and 24 and 22, respectively. A line joining the values of X_1 for the years 8 and 11 should indicate the relation between X_1 and X_2 with X_3 held constant at 9, and X_4 at 22 to 24. This line connecting these points was drawn on figure 14 and labeled *A*.

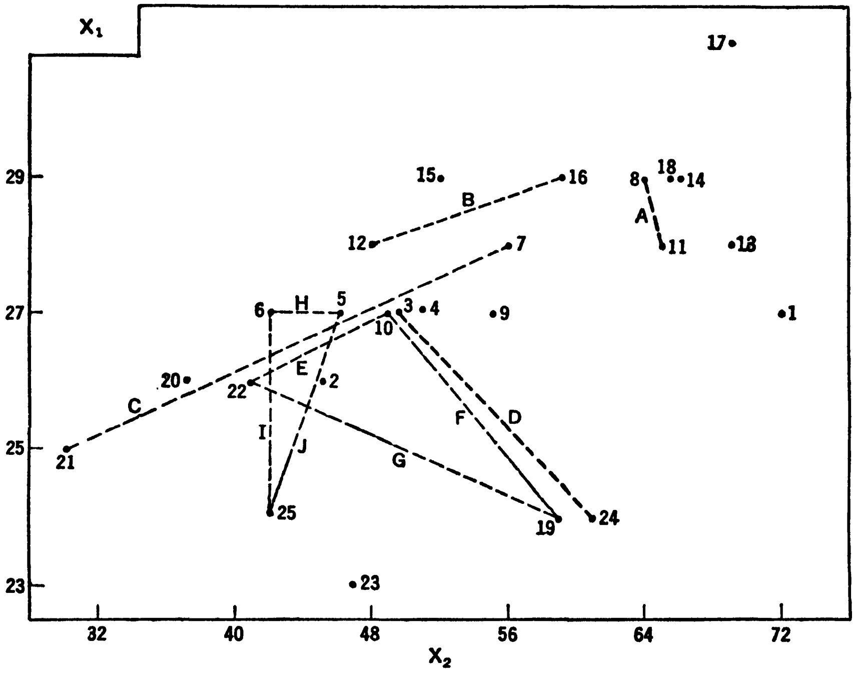


FIGURE 14.—GUIDES FOR DRAWING THE X_1X_2 FIRST APPROXIMATION CURVE

The 10 lines, *A* to *J*, connect points for which the values of X_3 and X_4 were the same.

Since these 10 lines had no uniformity of direction, they would have been of little aid in establishing the curve of relationship shown in figure 8.

For the years 12 and 16, the values of X_3, 12 and 13, and X_4, 23 and 20, were approximately equal. A line joining the values of X_1 for these two years should indicate the relation between X_1 and X_2 holding X_3 constant at 12 to 13 and X_4 at 20 to 23. The line connecting the two points was labeled *B* in figure 14. A line joining the 2 years 7 and 21 should indicate the relation of X_1 and X_2 with X_3 held constant at 11 and X_4 at 22 to 24 (figure 14, line *C*). Similarly, a line could be drawn between years 3 and 24 (figure 14, line *D*).

Three lines joining the value of X_1 for the years 10, 19, and 22 should indicate the relationship between X_1 and X_2 holding constant X_3 at 15 to 16 and X_4 at 23 to 24, labeled *E*, *F*, and *G*. Similarly, a line could be drawn through the years 5, 6, and 25 (figure 14, labeled *H*, *I*, *J*). From all these lines, it is assumed that the student will be able to determine the relation between X_1 and X_2 for all values of X_2. Three lines, *B*, *C*, and *E*, suggested that the acreage of corn increases moderately with the price of corn; one line, *J*, indicated a sharp increase; three lines, *D*, *F*, and *G*, suggested a moderate decrease; one, *A*, a rapid decrease; and *H* and *I* showed no relationship. By examining the lines alone, the student would be in a quandary as to where to draw a curve of relationship. By looking at both the lines and all the points, the student might be able to draw the curve. It is doubtful whether the lines would be of much assistance in addition to the individual points. In this particular example, the student could probably observe the relationships more clearly without the lines than with them (compare figures 8 and 14).

This guide is especially helpful where the interrelationships are important and the index of correlation is very high. Under such conditions, the use of the guide will establish the correct curves with fewer approximations than otherwise would be needed.[28] When the index of correlation is not very high, regardless of whether interrelationships are present, the guide is more likely to confuse than to aid.

CHARACTERISTICS OF CURVILINEAR METHODS

The three curves for the X_1X_2, X_1X_3, and X_1X_4 relationships were much the same for the least-squares and the two approximation methods. As the price of corn, X_2, fell, the acreage of corn planted decreased at an increasing rate. Of the three methods, the least-squares curve departed least from linearity (compare figure 1 with figures 5 and 11). The two approximation curves were about the same (compare figures 5 and 11).

As the price of cotton, X_3, increased, the acreage of corn, X_1, decreased at an increasing rate. Again, the least-squares curve departed least from linearity, and the more curvilinear approximations were about the same (figures 1, 6, and 12).

There was doubt as to the significance of the X_1X_4 relationship. It appeared that an increase in stocks of corn, X_4, was accompanied by an approximately constant increase in acreage. The curves were linear in the least-squares and short-cut methods (figures 1 and 13). The Ezekiel approximation was a curve with a slight bend (figure 7).

---

[28] The device may be useful in other ways. It may reveal that the relationships considered are not additive, but multiplicative or joint. Its value in these preliminary investigations may exceed its value in curve drawing.

The indexes of multiple correlation were: least squares, 0.701; Ezekiel, 0.746; and short-cut, 0.768 (tables 1, 4, and 5). The two approximation methods yielded about the same results. The index by the least-squares method was less because the curves were less flexible.

The procedures in the three curvilinear methods have some fundamental differences. In the least-squares method, the shapes of the curves of relationship are assumed at the outset. The flexibility[29] of the curves is limited by their mathematical definition. In the approximation methods, the shapes of the curves are not assumed at the beginning, but are determined from the data as the work progresses.

In the least-squares and Ezekiel methods, multiple correlation is used, while in the short-cut method, it is not. Consequently, the short-cut method involves much less work.

In least-squares and Ezekiel methods, the effects of independent variables are *considered simultaneously*. This is accomplished by using multiple correlation analysis. Later in the Ezekiel method, each additional approximation for each curve is made *independently* of the approximations for the other curves. In the short-cut method, the independent variables are *not considered simultaneously*. They are considered *successively* in their order of importance. Throughout the short-cut process, the unexplained variability in one relationship is always related to the next independent variable.

The curves from all three methods supposedly take into account interrelationships among the independent variables. In the least-squares and Ezekiel methods, this is done in multiple correlation. In the short-cut method, it is attempted by (*a*) guides to drawing the curves so that the effects of the factors not considered in the relation will be held constant; (*b*) treating the independent variables in the order of their importance; and (*c*) making two or more sets of approximations.

The three methods differ in the nature and amount of personal judgment. In the least-squares methods, judgment determines only the types of curves to use. The mathematical procedure determines the positions of the curves. In the Ezekiel and short-cut methods, personal judgment determines both the shape and the location of the curves. The Ezekiel method requires less judgment than the short-cut method because the linear net relationships are determined mathematically and give some clues to the general direction of curves. In the short-cut method, the entire procedure is based on judgment.

---

[29] For example, the curve $Y = a + b \log X$ can take two general shapes: (*a*) increasing at a decreasing rate, or (*b*) decreasing at a decreasing rate; but the rate of increase or decrease is rigidly proportional to the logarithm of the independent variable.

The advantages of the least-squares method are as follows:

1. The location of the curves is determined mathematically.
2. Interrelationships between independent variables are considered automatically, and the resulting curves describe the net relationships.
3. The curves may be described by regression coefficients and regression equations.

The disadvantages of the least-squares method are as follows:

1. The most suitable curve for the relationship cannot always be expressed simply in mathematical terms. The method of least squares cannot be simply applied to all mathematical curves.
2. The shape of the curve must be assumed at the outset. This involves personal judgment. Even the experienced worker may not choose the correct curves, because he does not know the exact nature of the relationships at this stage of the analysis.
3. Because of errors in the assumption of the nature of the relationships and faulty judgment in choosing the correct mathematical expression of the curves, the amount of work required may be excessive.

The advantages of the Ezekiel approximation method are as follows:

1. The shapes of the curves are not assumed at the beginning, but are determined by the analysis.
2. The worker is guided in drawing the curves by the results of multiple correlation analysis. The multiple regression equation assumes linear relations, but the direction of these relations is usually the same as for the final curves. Moreover, in these linear relations, the interrelationships among independent variables have been considered.
3. Proceeding from the multiple correlation part of the method, each new curve is drawn independently of the other curves in that set of approximations.

The disadvantages of the Ezekiel approximation method are as follows:

1. Some personal judgment is required in plotting points, drawing the approximations, and reading the values from those curves.
2. It requires a large amount of work—more than either the least-squares or short-cut methods.
3. Since the curves are not expressed mathematically, their reliability cannot be accurately tested. There is a tendency on the part of

research workers to place too much reliance on the results of approximation analysis.

The advantages of the short-cut approximation method are as follows:

1. It involves relatively little work.
2. The shapes of the curves are not assumed at the outset.
3. It is well adapted to preliminary analysis.

The disadvantages of the short-cut method are as follows:

1. The analysis from beginning to end involves much personal judgment.
2. It is doubtful whether this method considers the effect of interrelationships as accurately as the other methods. This is especially true when the ρ is not high.
3. Too much reliability is sometimes placed on the curves. The inductive value of the curves is difficult to test accurately.[30]

USES

Multiple curvilinear correlation analysis has been widely used in many fields of scientific research. In general, its applications are the same as those for linear analysis, except for the nature of the relationships described.

Campbell[31] used the least-squares method of determining the price of rice from 1914 to 1930, X_1, in terms of the United States supply, X_2, and Indian production, X_3. As the United States supply increased, the price decreased at a decreasing rate. As Indian production increased, the United States price declined at a uniform rate. The index of correlation was 0.985.

Elliott[32] used the Ezekiel method of approximating the curvilinear relationship existing between the September to April receipts of hogs at Chicago from 1898 to 1916, X_1, and the corn-hog ratio for December, X_2, for June to November, X_3, for January to March, X_4, the climate at farrowing time, X_5, and long-time trend, X_6. The index of correlation was 0.983. The corn-hog ratio accounted for 72 per cent of the variation in hog receipts; climate, 18 per cent; and trend, about 7 per cent.

---

[30] A detailed discussion of advantages and disadvantages of the short-cut method was given by Malenbaum, W., and Black, J. D., The Use of the Short-Cut Graphic Method of Multiple Correlation, Quarterly Journal of Economics, Volume LII, No. 1, pp. 66–112, November 1937; and Bean, L. H., Ezekiel, M., Black, J. D., and Malenbaum, W., Comments, Rejoinder, and Remarks on the Short-Cut Graphic Method of Multiple Correlation, Quarterly Journal of Economics, Volume LIV, No. 2, pp. 318–364, February 1940.

[31] Campbell, C. E., Factors Affecting the Price of Rice, United States Department of Agriculture, Technical Bulletin No. 297, pp. 21–23, April 1932. The author also used the Ezekiel approximation method on other aspects of the price of rice, page 31.

[32] Elliott, F. F., Adjusting Hog Production to Market Demand, University of Illinois Agricultural Experiment Station, Bulletin 293, pp. 557–560, June 1927.

Ratcliffe[33] used the short-cut method to analyze the monthly Minneapolis price of flaxseed, X_1, from 1922 to 1931. He related the 10 October prices to the index of prices of all commodities, X_2, the Argentine supply, X_3, and the Argentine new crop estimate for October, X_4. The price of flaxseed increased at an increasing rate with the index of all commodities; decreased at an increasing rate with the Argentine supply; and decreased at a decreasing rate with the estimate of Argentine production. The index of multiple correlation was 0.975.

---

[33] Ratcliffe, H. E., Flaxseed, North Dakota Agricultural Experimental Station, Bulletin 268—Technical, pp. 10–37, February, 1933.

CHAPTER 14

JOINT CORRELATION

All the multiple correlation methods treated thus far have one common fault. They assume that the relationships between each independent and the dependent variable are additive.[1] In other words, the effect of one independent variable has been assumed to be constant for all values of the other independent variables. In the typical linear multiple regression equation, $X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$, the effect of a given change in X_2 on the size of X_1 is constant, regardless of the size of X_3. In this additive relationship, the effect of X_2 on X_1 is independent of X_3.

When a relationship is not additive, it is joint. In a joint relationship, the effect of X_2 on X_1 is dependent on X_3. That is, X_2 may have a greater effect on X_1 when X_3 is large than when X_3 is small.

Like additive relationships, treated under the subjects of linear and curvilinear multiple correlation, joint relationships may also be either linear or curvilinear. In linear joint analysis, the change in the effect of X_2 with different values of X_3 would be a change in the rate of change in X_1 in terms of X_2. Since most relationships are not linear, the change in the effect of X_2 may take the form of changing not only the slope of the curve, but also the very nature of the relationship as shown by the shape of the curve.

Joint relationship can be analyzed by either least-squares methods or by approximation methods.

LEAST-SQUARES METHOD

The least-squares method may be used to analyze either linear or curvilinear joint relationships.

LINEAR JOINT CORRELATION

Profits from growing apples, X_1, are associated with size of orchard, X_2, and the yield, X_3. For New York fruit farms, the multiple correlation coefficient was $R_{1.23} = 0.822$. The linear regression equation was $X_1 = 30.9X_2 + 25.7X_3 - 2{,}865$ (table 1). Each additional acre of orchard, X_2, and each bushel per acre, X_3, added $31 and $26, respectively, to

---

[1] Additive and joint relationships were discussed briefly on pages 128 to 134.

TABLE 1.—LINEAR JOINT CORRELATION*

20 NEW YORK FRUIT FARMS, 1936

X_1 = Profit ÷ 500; X_2 = Acres of orchard ÷ 10;
X_3 = Bushels per acre ÷ 20; $X_4 = X_2X_3$

X_1	X_2	X_3	X_2X_3 represented by X_4
19	9	12	108
3	9	9	81
7	8	8	64
1	7	3	21
6	6	10	60
10	6	9	54
13	5	11	55
−2	4	4	16
9	4	12	48
0	3	4	12
3	3	5	15
2	3	10	30
−1	3	6	18
6	2	10	20
0	2	3	6
−1	2	6	12
2	2	5	10
0	2	2	4
−1	1	1	1
1	1	7	7
Total 77	82	137	642
Average 3.85	4.1	6.85	32.1

Linear joint relationship

$$X_1 = -48.5X_2 + 6.48X_3 + 0.510X_2X_3 - 253$$

$\rho^2_{1.23\,(\text{linear joint})} = 0.759$ $\rho = 0.871$

$\rho_{1.23\,(\text{linear joint})} = R_{1.234}$

For comparative purposes, the *linear additive* multiple relationship is presented.

$$X_1 = +30.9X_2 + 25.7X_3 - 2{,}865$$

$R^2_{1.23} = 0.676$ $R_{1.23} = 0.822$

* Since the numbers in the three variables were large, they were coded as given in the subtitle to table 1. For instance, the profit from the first farm was $9,408. This number was divided by $500, and the quotient, 19, was the first item recorded in the first column of table 1 under X_1.

The calculations of the products X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, and X_3X_4; the squares X_1^2, X_2^2, X_3^2, and X_4^2; the product moments; standard deviations; the solution of the normal equations; and the like have been omitted. They are exactly the same as for four variables given in the chapter on multiple correlation, pages 171 to 174.

The student should keep in mind that X_4 stands for the product X_2X_3. Therefore, the products $X_1X_4 = X_1X_2X_3$; $X_2X_4 = X_2^2X_3$; and $X_3X_4 = X_2X_3^2$. The square $X_4^2 = (X_2X_3)^2 = X_2^2X_3^2$.

After the solution of the normal equations, the regression equation in terms of coded data was

$$X_1 = -0.970X_2 + 0.259X_3 + 0.204X_4 - 0.505$$

and in terms of original values,

$$X_1 = -0.970\,\frac{(500)}{(10)}\,X_2 + 0.259\,\frac{(500)}{(20)}\,X_3 + \frac{0.204(500)}{(10)\,(20)}\,X_2X_3 - 0.505(500)$$

$$X_1 = -48.5X_2 + 6.48X_3 + 0.510X_2X_3 - 253$$

the profit. According to the equation, an additional acre, X_2, added \$31 profit, regardless of whether the yield, X_3, was high or low. Likewise, each additional bushel, X_3, added about \$26 profit, regardless of whether the orchard was large or small. The principle stated in the equation could be challenged as not in accordance with fact. It may be the farmer's experience that, though a large farm will make more profit than a small one when yields are high, it will make less, or lose more, than the small farm when yields are poor. A large yield would probably not add so much to the profit on a small farm as on a large one. If the relationship is of this nature, it is joint. A joint relationship cannot be shown by a regression equation of the type $X_1 = 30.9X_2 + 25.7X_3 - 2{,}865$.

One of the simplest expressions of a joint relationship is an equation of the type $X_1 = a + bX_2 + cX_3 + dX_2X_3$. The joint relationship is expressed by the product of the two independent variables, X_2X_3. This equation may be called a *linear* joint regression equation, because, when X_3 is held constant, the expression reduces to a linear relationship between X_1 and X_2. If X_3 is held constant at several different levels, the resulting relationships between X_1 and X_2 will all be linear, but the straight lines will have different slopes. When X_2 is held constant, the same principles hold for the X_1X_3 relationship.

The simplest method of handling the joint relationship X_2X_3 is to call it a fourth variable, X_4. Then, the usual linear multiple correlation procedure for four variables is followed. For profits on fruit farms, such a regression equation was $X_1 = -48.5X_2 + 6.48X_3 + 0.510X_4 - 253$. Since $X_4 = X_2X_3$, the regression equation was really $X_1 = -48.5X_2 + 6.48X_3 + 0.510X_2X_3 - 253$. The individual parts of this equation considered separately are practically meaningless. One of the first three terms cannot be held constant without holding one of the others constant. If two of these terms are held constant, the remaining term is automatically a constant. The meaning of the equation becomes clear when fixed values of one variable, say X_3, are assumed. For instance, when yield, X_3, was 50 bushels, the equation read:

$$X_1 = -48.5X_2 + 6.48\,(50) + 0.510(50X_2) - 253$$
$$X_1 = -48.5X_2 + 324 + 25.50X_2 - 253$$
$$X_1 = -23.0X_2 + 71$$

This states that, when the yield was 50 bushels, each additional acre of orchard *reduced* profits by \$23. With yields of 125 bushels, the joint equation reduces to $X_1 = +15.3X_2 + 557$; and with yields of 200 bushels, to $X_1 = +53.5X_2 + 1{,}043$. The joint relationship stated that size of orchards had a negative effect when yields were low, $-\$23$;

little or no effect when yields were moderate, +\$15; and considerable positive effect when high, +\$54. The joint regression equation permitted size of orchard, X_2, to have a changing effect on profits, X_1, with different yields, X_3.

In the joint analysis, each of the X_1X_2 relationships was linear, but none were the same as the relationship found with linear additive analysis ($X_1 = +30.9X_2 + \cdots$ lower right, table 1). In this linear additive relationship, the regression coefficient measuring the effect of the size of orchard, X_2, was assumed to be always the same, +\$31, regardless of yields, X_3.

TABLE 2.—COMPARISON OF ADDITIVE AND JOINT RELATIONSHIPS

EFFECT OF SIZE OF ORCHARD, X_2, AND YIELDS PER ACRE, X_3, ON PROFIT, X_1

Size of orchard, acres, X_2	Linear additive—yields per acre, X_3			Linear joint—yields per acre, X_3			Change per bushel with a given acreage	
	50 bu.	125 bu.	200 bu.	50 bu.	125 bu.	200 bu.	Linear	Joint
	Profits, X_1	*Profits*, X_1	*Profits*, X_1	*Profits*, X_1	*Profits*, X_1	*Profits*, X_1	*Profits*, X_1	*Profits*, X_1
10	\$ −1,271	\$ 657	\$2,584	\$ − 159	\$ 710	\$1,578	\$ *+26*	\$ *+12*
40	− 344	1,584	3,511	− 849	1,167	3,183	*+26*	*+27*
70	583	2,511	4,438	−1,539	1,625	4,788	*+26*	*+42*
Change per acre with a given yield.........	*+31*	*+31*	*+31*	*−23*	*+15*	*+54*		

The differences between additive and joint relationships are shown by the coefficients in the above equations. The comparison of these additive and joint relationships may be simplified by calculating estimated profits for different combinations of size of orchard and yields (table 2). Reading *down* the columns of table 2 is equivalent to examining the effect of acreage, X_2, on profits, X_1, with yields, X_3, held constant. With the additive analysis, profits *increased* from −\$1,271 to +\$583, as the size of orchard increased when yields were small. With the joint analysis, profits *decreased* from −\$159 to −\$1,539. Similar comparisons could be made for moderate and high yields.

The joint relationship probably presented the truer picture (table 2). The amount of money a farmer makes or loses because of higher or lower yields does depend upon the size of the orchard. However, the linear relationship states that, even though the crop is a failure, a farmer can profit if he has a large enough orchard. Experience has proved that, with a poor crop, the large orchard loses more, that is, makes less, than the small orchard.

Thus far, only the effect of acreage on profits has been examined. The effect of yield on profits for orchards of different sizes is also a joint relationship. It is almost self-evident that, if X_2 is jointly related with X_3, then X_3 is jointly related with X_2.

The effect of yields, X_3, on profits, X_1, can be examined by reading *across* the rows of table 2. According to the linear additive relationship, an increase of 1 bushel per acre in yield always increased profits by $26 (table 2). The joint relationship stated that, though each additional bushel per acre increased profits by $42 for a 70-acre orchard, it increased profits by only $12 for a 10-acre orchard. The joint relationship seems to agree with experience. With a small orchard, a farmer does not make so much because of high or low yields as with a large orchard.

It is possible to calculate an index of joint correlation. In linear joint analysis, the index is the same as the multiple correlation coefficient, $\rho_{1.23(\text{linear joint})} = R_{1.234} = 0.871$, where X_4 is X_2X_3.

A comparison of the coefficients of determination, $R^2_{1.23} = 0.676$ for the additive analysis, and $\rho^2_{1.23} = 0.759$ for the joint analysis, indicated that the joint relationship was probably more accurate than the additive relationship[2] (table 1).

CURVILINEAR JOINT CORRELATION

Most relationships, whether joint or additive, are not linear. In practice, the relatively simple linear joint regression equation is only rarely applicable. It is possible to show simpler types of non-linear joint correlation with mathematical curves.

The relations of rainfall and temperature to crop yields are usually not linear and not additive. For example, the relationships of rainfall, X_3, and temperature, X_2, to the yield of corn in six leading states, X_1, are curvilinear and joint. With the methods used in linear joint analysis (table 1), the following curvilinear joint expression was calculated:

$$X_1 = 57.66X_2 + 2.31X_3 - 0.28X_2^2 - 0.004X_3^2 - 0.012X_2X_3 - 2{,}931.96$$

This equation contains the following six variables:

1st variable, X_1 = yield of corn
2nd variable, X_2 = June temperature
3d variable, X_3 = July-August rainfall

4th variable, X_2^2 = temperature squared
5th variable, X_3^2 = rainfall squared
6th variable, X_2X_3 = temperature times rainfall

The index of curvilinear joint correlation was $\rho_{1.23(\text{curvilinear joint})} = 0.673$.

---

[2] The significance of the difference between $R^2_{1.23}$ and $\rho^2_{1.23\,(\text{linear joint})}$ is discussed on pages 417 to 419.

The terms in X_2 and X_2^2 describe a curvilinear relationship between X_1 and X_2. As temperature rises, yields first increase and then decrease.

The terms in X_3 and X_3^2 describe a curvilinear relationship between X_1 and X_3. As rainfall increases, yields first increase and then decrease.

The term X_2X_3 changes the nature of the two curvilinear relationships with different combinations of X_2 and X_3. In common parlance, this term states that effects of both temperature and rainfall on yield are not the same for different amounts of either. It further states that the effect of rainfall is not the same when temperature is high as when low; and that the effect of temperature is not the same when rainfall varies. The nature of the relationship can be further examined from the estimated yields of corn for different combinations of temperature and rainfall presented in tabular form (table 3).

TABLE 3.—RELATION OF JUNE TEMPERATURE, X_2, AND JULY-AUGUST RAINFALL, X_3, TO YIELD OF CORN IN SIX LEADING STATES*

ANALYSIS OF CURVILINEAR JOINT RELATIONSHIPS BY LEAST-SQUARES METHOD
THE VARIABLES WERE EXPRESSED IN PERCENTAGE OF NORMAL

Rainfall X_3	Temperature, X_2			
	94	98	102	106
	Yield corn	*Yield corn*	*Yield corn*	*Yield corn*
60	71	83	87	82
90	88	99	102	95
120	98	108	109	101
150	101	110	109	100

* Based on the regression equation

$$X_1 = 57.66X_2 + 2.31X_3 - 0.28X_2^2 - 0.004X_3^2 - 0.012X_2X_3 - 2,931.96$$

The yield of corn varied from 71 to 110 per cent of normal.

With temperature, X_2, constant at 94 per cent of normal, the increases of 30 in rainfall increased yields +17, +10, and +3. With temperature constant at 106, the changes in yields were +13, +6, and −1.

With rainfall constant at 60, the increases of 4 in temperature changed yields +12, +4, and −5. With rainfall constant at 50 per cent above normal, the changes in yields were +9, −1, and −9.

When rainfall was 60, the highest yield was obtained when temperature was 102; and when rainfall was 150, the high yields came with temperatures of 98.

The possibilities of mathematical curves in joint correlation are limited. One difficulty is that the simpler ones are not sufficiently flexible to describe the relationships; and the more complicated expressions are difficult and often impractical to fit. The most important inadequacy of mathematical expressions lies in the inability to predetermine exactly which expression will best describe the data. The number of possible joint regression equations is almost unlimited. In the two joint regression equations given thus far in this chapter, the interaction of X_2 and X_3 on X_1 has been shown as a product, X_2X_3. In practice, the interaction might be $\frac{X_2}{X_3}, \frac{X_3}{X_2}, \frac{X_2^2}{X_3}, \frac{\log X_2}{\sqrt{X_3}}, \frac{1}{X_2X_3}, \frac{X_2X_3}{X_2 - X_3}$, or other more complicated forms. At the beginning of a correlation problem, there is no way of determining which of the above expressions best fits the relationship at hand. It was pointed out in the discussion of additive relationships that the nature of curves was difficult to predetermine in multiple analysis (page 217). The nature of joint relationships is even more difficult to predetermine. The joint term of an equation does not describe the relationships between the independent and dependent variables, but describes the relationship between two relationships. The nature of the joint relation is usually so deeply buried that it is difficult to visualize, describe, or measure.

APPROXIMATION METHOD

Probably the most common method of analyzing joint relationships is the graphic approximation method in which the nature of the joint relationship is not assumed at the outset. With this method, the nature of the relationships unfolds as the work progresses.

Plotting Three Variables on a Two-Dimensional Graph

The first step is to plot the observations on a graph. The horizontal and vertical scales measure the two independent variables X_3 and X_2. Thus, the location of each observation on the chart is determined by the size of X_3 and X_2. The value of each point on the chart is simply the value of X_1. During 1890, rainfall, X_3, was below normal, 83; the temperature, X_2, was in excess of normal, 105; and the yield, X_1, was 90. The point for 1890 was placed at 83 on the horizontal scale and at 105 on the vertical scale. The value for that point was 90, which was recorded at its exact location[3] (figure 1).

During 1891, rainfall, X_3, was more than normal, 108; temperature, X_2, was about normal, 99; and yield, X_1, was high, 123. The point on the chart was called 123 and was placed at 108 on the horizontal scale and at 99 on the vertical scale. The yields for the other 36 years were plotted in the same manner.

---

[3] The point was labeled "90" because the yield was 90, and not because the year was '90.

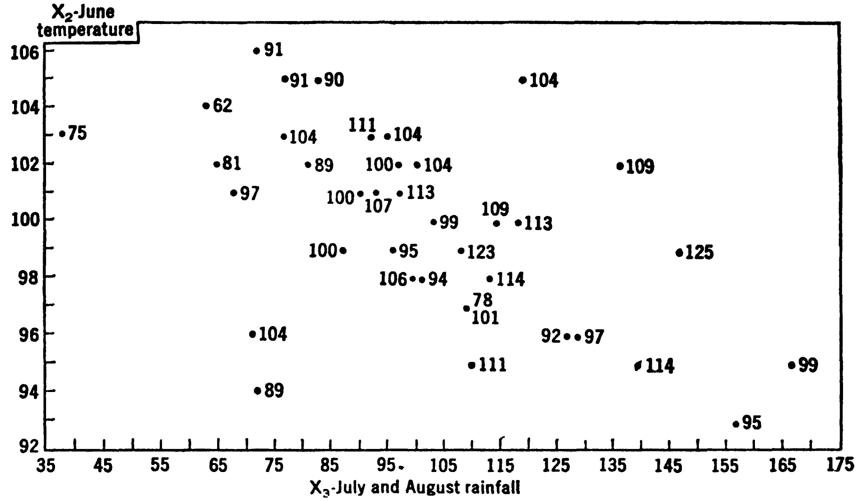


FIGURE 1.—THREE VARIABLES PLOTTED ON A TWO-DIMENSIONAL CHART

YIELDS OF CORN PLOTTED WITH TEMPERATURE AND RAINFALL

The independent variables, temperature and rainfall, were measured along the vertical and horizontal scales, respectively. The dependent variable, yield of corn, is indicated by the size of the numbers accompanying the points.

This chart is not similar to any previously used in showing relationships. In the study of other correlation coefficients and indexes, the charts always showed the relationship between the independent and the dependent variables. The vertical scale was the dependent, and the horizontal scale the independent variable. The plotted points were labeled with some number for *identification*, but they had no values themselves and did not measure any variable. In the present problem, both the vertical and horizontal scales measure independent variables. The dependent variable is not measured by a given scale but by the size of the numbers accompanying the plotted points. This chart may be thought of as having three dimensions, width, length, and height. Temperature, X_2, on the vertical scale may be called width. Rainfall, X_3, on the horizontal scale may be called length. The yield, X_1, is height, the third dimension, and is not shown graphically. However, this third dimension is indicated by the size of the numbers which accompany the plotted points.

An analogy to this type of relationship among three variables is the physical features of an area of land. One independent variable, X_2, is represented by the distance north and south; the second, X_3, by the distance east and west. The dependent variable, X_1, is represented by the altitude of the land, hills, valleys, plateaus, bluffs, and the like. A common map directly shows the length and breadth of land graphically. The topography or altitude of the earth's surface is indicated in one of two ways. The distance above sea level may be written on the map at the location in question. This is exactly the same method of indicating the third dimension that was employed for yields of corn, figure 1.

DRAWING CONTOUR LINES

The common method of showing topography on a map is drawing contour lines. A contour is a line drawn connecting all the points of a given elevation, say 500 feet. For every point on the contour line, the value of the third dimension, elevation above sea level, is supposedly the same, regardless of the distance from north to south or east to west. The accuracy of the contour lines depends upon the number of bench marks, the particular surveyor, and the time he spent on the area.

The relationship of temperature and rainfall to yield may be generalized with the use of contour lines. If the relationship were perfect, contour lines could be constructed in figure 1 by connecting all points with the same values. For example, all points with values of 90 could be connected by one line or curve; and all points with values of 100 might also fall on one contour line. However, in this problem, the relationship is not perfect; consequently, the positions of the contour lines must be approximated.

The first task is to study very carefully the values of these points relative to their location (figure 1). The two highest yields, 123 and 125, occurred during years of average or more rainfall and with temperature slightly below normal. Most yields of 5 per cent or more above normal occurred when July and August rainfall, X_3, was average or above, and when the temperature, X_2, was cool rather than hot (figure 1). Most low yields occurred when rainfall was light and temperature was above normal. However, there were several exceptions to these generalizations.[4]

In the drawing of the contour lines, the student must keep in mind that the lines should fit the points as closely as possible. However, the lines should conform to some general patterns and show definite principles. The lines should, of course, not cross one another and should not "wiggle" about indiscriminately. No contour should depart too far from those on either side of it, and all contour lines should be long, sweeping curves or straight lines.

---

[4] In 1924, temperature was below normal, $X_2 = 97$, and rainfall was somewhat above normal, $X_3 = 109$. With these conditions, a good yield would be expected, but it was low, $X_1 = 78$. This was due to factors not considered in this analysis. The spring was late and wet; there was an early frost; and much of the unusually late crop failed to mature.

It must be remembered, however, that, the smaller the differences between the values of the points and of the contours passing through the points, the better the fit. The problem is to obtain as close a fit as possible and still have a sensible and conservative set of contours. In accordance with the above principles, a set of contour lines was drawn on figure 1 (shown in figure 2). Each contour line is labeled with a value of X_1, which is the estimated value of X_1 for all points on the contour.

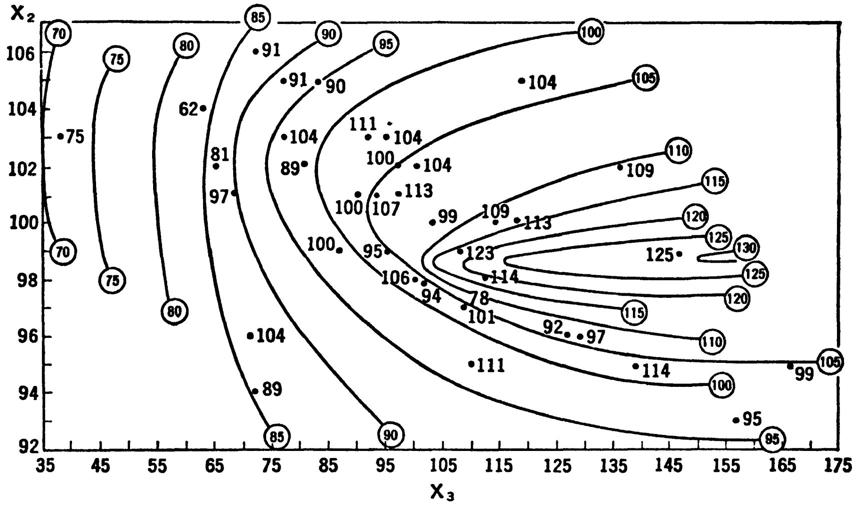


FIGURE 2.—CONTOUR LINES DESCRIBING THE JOINT RELATIONSHIP BETWEEN YIELDS OF CORN AND TEMPERATURE AND RAINFALL

The contour lines are the estimated yields for various combinations of June temperature and July and August rainfall.

ESTIMATED X_1, RESIDUALS, AND RHO

The estimated value for any point in figure 2 may be determined by interpolating between two contour lines. For instance, in 1892, the actual yield was 100. This point is to the left of the center of figure 2, between the 100 and the 105 contour lines. From the contours, the value of X_1 was estimated to be 104, as it was much closer to the 105 than to the 100 contour. Similarly, in 1924, the actual yield, 78, fell between the 100 and the 105 contour lines, and the estimated yield was

104. These estimated values, X'_1, read from the contours, were recorded in table 4.

TABLE 4.—INDEX OF JOINT CORRELATION APPROXIMATED FROM CONTOUR LINES

JUNE TEMPERATURE AND JULY AND AUGUST RAINFALL RELATED TO THE YIELD OF CORN, 1890–1927

Year	Temperature X_2	Rainfall X_3	Actual yield X_1	Estimated yield from contours X'_1	Residuals $X_1 - X'_1$ z	z^2
1890	105	83	90	95	− 5	25
1891	99	108	123	117	+ 6	36
1892	101	90	100	104	− 4	16
1893	101	68	97	89	+ 8	64
1894	103	38	75	72	+ 3	9
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1924	97	109	78	104	−26	676
1925	103	92	111	102	+ 9	81
1926	96	127	92	106	−14	196
1927	99	96	95	105	−10	100
Total	3,795	3,820	3,790	—	—	3,024
Average	99.87	100.53	99.74	—	—	79.58

$$\rho_{(\text{approximation joint, contours})} = \sqrt{1 - \frac{79.58}{158.93^*}} = \sqrt{1 - 0.5007} = \sqrt{0.4993}$$

$$= 0.707$$

* $\Sigma X_1^2 = 384{,}042$; and $\sigma_1^2 = 158.93$.

From this point, the calculation of the index of correlation is simple, proceeding as usual from the squared residuals:

$$\rho_{1.23(\text{curvilinear joint, contours})} = \sqrt{1 - \frac{S_{1.23}^2}{\sigma_1^2}} = 0.707$$

This is a measure of the degree of the relationship between climate and yield shown in figure 2. The coefficient of determination, $\rho^2 = 0.50$, measures the proportion of the squared variability in yield associated with differences in June temperature and the July and August rainfall. Probably the temperature and rainfall in other months of the growing season explained a considerable part of the unaccounted-for squared variability, 0.50.

The value of rho from contour lines of joint relationship varies considerably with the particular set of contours drawn. The difference between $\rho_{1.23(\text{curvilinear joint, contours})} = 0.707$ and the true degree of relationship depends upon how well the contours were drawn. If the contours were drawn in the correct positions except for minor "wiggles" in the curves, $\rho_{1.23} = 0.707$ is probably too high. If the contours were sufficiently conservative and sensible but were drawn in the wrong positions, $\rho_{1.23} = 0.707$ is probably too low.

As in other types of approximation correlation analysis, the work may be repeated in an attempt to improve the relationship. An examination of the residuals from the first set of contours may give some clue for improvement. The lines might be revised so that the extremely large residuals were decreased, even though the revisions increased some of the smaller residuals.

TABLE 5.—TABULAR SUMMARY OF THE JOINT RELATIONSHIP OF TEMPERATURE AND RAINFALL TO YIELD

ESTIMATED FROM CONTOUR LINES IN FIGURE 2

ANALYSIS OF CURVILINEAR JOINT RELATIONSHIPS BY APPROXIMATION METHOD

Rainfall X_3	Temperature, X_2				
	94	97	100	103	106
	Yield, X'_1	*Yield*, X'_1	*Yield*, X'_1	*Yield*, X'_1	*Yield*, X'_1
70	84	87	90	91	86
85	89	93	100	101	92
100	93	99	108	104	98
115	96	108	113	107	101
130	98	113	118	108	102

GRAPHIC AND TABULAR METHODS OF SHOWING JOINT RELATIONSHIPS

It is difficult to visualize a joint relationship. It is almost impossible to draw a graph which accurately describes a joint relationship and is simple and easy to understand. A contour chart such as figure 2 tells the whole story, but is very difficult to interpret. Each dimension represented one factor. Since there were two dimensions and three factors, only two factors could be shown graphically. The other factor is shown numerically. The drawing of contours clarifies the picture but little.

The ideal graph for showing joint relationships between two independent and one dependent variable is three-dimensional. The simplest procedure is first to construct a table of the estimated yields, X'_1, for varying combinations of temperature and rainfall. These values can be read directly with the aid of the contours in figure 2. Five different amounts of rainfall and of temperature were selected, giving 25 different combinations (table 5). With a low rainfall, $X_3 = 70$, and cool weather, $X_2 = 94$, the point in figure 2 lay between the 80 and 85 contour lines at the lower left of figure 2. The value of X_1 was estimated at 84. The same procedure was followed in estimating the other 24 values in table 5. These values represent the third dimension. The next problem is to superimpose this third dimension on a plane representing the other two dimensions. This involves the construction of a box with vertical sides of variable height and a top with a variable shape.

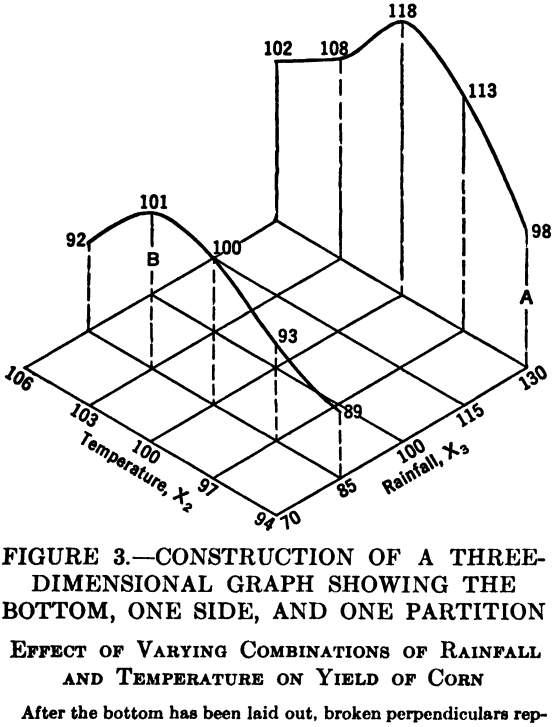


FIGURE 3.—CONSTRUCTION OF A THREE-DIMENSIONAL GRAPH SHOWING THE BOTTOM, ONE SIDE, AND ONE PARTITION

EFFECT OF VARYING COMBINATIONS OF RAINFALL AND TEMPERATURE ON YIELD OF CORN

After the bottom has been laid out, broken perpendiculars representing the estimated yield of corn, *A* and *B*, are erected from points representing varying combinations of rainfall and temperature.[6]

The bottom of the box is usually made first. In the corn-yield problem, the bottom was square. One side of the square represented a range of 70 to 130 in rainfall, X_3; and the other side, a range of 94 to 106 in temperature, X_2. An oblique view of this bottom gives the erroneous impression that it is diamond-shaped rather than square (figure 3).

The heights of the box for different combinations of rain and yield are the values of X_1 shown in table 5. With a rainfall of 130, the yields with varying temperature were 98, 113, 118, 108, and 102 (table 5). When rainfall, X_3, was 130, and temperature, X_2, was 94, estimated yield, X_1, was 98. At the intersection of the two lines representing $X_3 = 130$ and $X_2 = 94$, a perpendicular, A, was erected representing 98. The same procedure was followed for the other estimated values of X_1 for rainfall of 130. When the points were connected, one side of the box, with a height varying from 98 to 118, had been constructed. The same procedure was followed in the constructing of the other three sides of the box and its six partitions. One of these partitions is shown in figure 3. Where temperature, $X_2 = 103$, and rainfall, $X_3 = 85$, intersect, a perpendicular, B, was erected representing yield, $X_1' = 101$. After all the sides and partitions have been constructed and the tops of the perpendiculars connected in both directions, it can be seen that these points determine an irregular surface (figure 4).

---

[6] For convenience, the bottom of the graph was placed at $X'_1 = 80$. If the bottom had been placed at $X'_1 = 0$, the box would have been higher, but the shapes of the tops of the sides and partitions would have been the same.

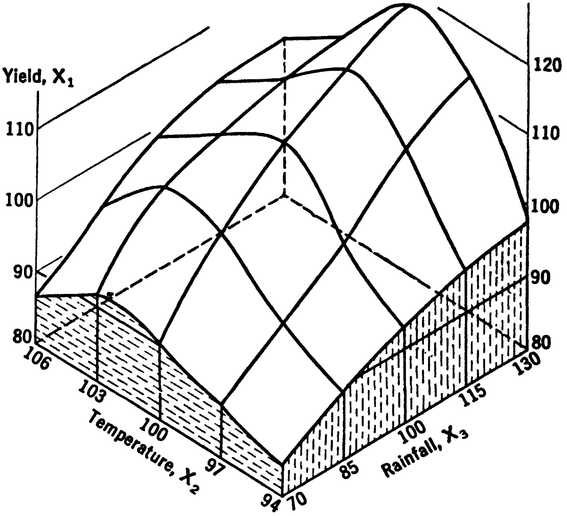


FIGURE 4.—THE SURFACE OF A THREE-DIMENSIONAL GRAPH

EFFECT OF VARYING COMBINATIONS OF RAINFALL AND TEMPERATURE ON YIELD OF CORN

The surface indicated by the curved lines shows changes in yield with varying combinations of temperature and rainfall.

The surface is an accurate description of the relationships involved. However, figure 4 contains three "invisible" broken lines to give a box effect, three scales, and ten important visible lines representing the surface. The human mind has difficulty in grasping the meaning of a chart with more than two or three lines. Consequently, such three-dimensional charts are more impressive than informative.

With cardboard or modeling clay, a solid form representing these relationships can be made. Such a model portrays the relationship more effectively than figure 4. This is merely stating that a three-dimensional graph can be shown more effectively in three than in two dimensions.

The most effective way of showing joint relationships is the simple tabular form (table 5).

JOINT CORRELATION WITH MORE THAN TWO INDEPENDENT VARIABLES

Joint correlation analysis is usually limited to two independent variables.

If there are several independent variables and only two are jointly related, the joint effect of these two on the dependent variable may be determined. First, the additive effects of all the independent variables could be eliminated.[6] With the residuals from the additive relationship as the dependent variable, the analysis proceeds in the manner presented in this chapter.

When several pairs of the independent variables have joint effects, but the effects of one pair are not associated with the effects of any other variables, the joint relationships may be shown by two or more three-dimensional graphs.

When three or more independent variables have joint effects, the problem becomes much more complicated. The nature of the relationship is almost impossible to determine mathematically. It is also impossible to show such a relationship graphically in four dimensions. Even if the methods of determining and showing such relationships were simple, the number of observations required to obtain reliable results would make the method unwieldy. With the necessary number of observations, the best way to analyze such relationships probably would be cross tabulation.[7]

LEAST-SQUARES *vs.* APPROXIMATION METHODS

Joint relations can be analyzed with least-squares equations or can be approximated graphically. The least-squares analysis has the advantage of being more mechanical and requiring less judgment. The relationship shown by an equation is rigidly defined. The equation can be conveniently used to estimate values of the dependent variable under different combinations of conditions. Rho from least-squares analysis is rigidly defined, and its significance can be tested.

Least-squares equations have decided disadvantages. It is often difficult to choose an equation which fits the relationship, even when the relationships are known. It is even more difficult to predetermine what the relationships are. Even if it is assumed that the relationships are known and that a satisfactory equation has been chosen, the amount of work in fitting the equation to the data is often prohibitive.

The approximation method of analyzing joint relationships does not assume the nature of the relationships at the outset. The relationships are determined as the analysis progresses. The amount of work involved is less than that required for the least-squares analysis even when the simplest equations are used. Approximation analysis is more flexible than least-squares analysis. This is both an advantage and a disadvantage. It is an advantage in that it is possible to obtain the true relationship by approximation; it is a disadvantage in that there may be unwarranted irregularity in the approximations, and rho may be too high.

---

[6] Regression equation, page 176.

[7] Joint relationships in tabulations are discussed on pages 283 and 374.

The weather and corn-yield problem was analyzed algebraically and graphically.[8] Using parabolas to show curved relationships and a product to show the joint relationship, the authors obtained $\rho_{1.23(\text{curvilinear joint})} = 0.673$. The relationship was stated in equation form as follows:

$$X_1 = 57.66X_2 + 2.31X_3 - 0.28X_2^2 - 0.004X_3^2 - 0.012X_2X_3 - 2{,}932 \text{ (page 250)}$$

If a three-dimensional graph of this relationship had been constructed, the surface would have resembled that in figure 4 but would have been more regular. Even though the approximation method was more flexible, $\rho = 0.707$, it was only a little larger than $\rho = 0.673$ from least-squares analysis. Since the latter index is probably the more reliable, it might appear that the least-squares method was superior. However, the equation used was chosen on the basis of the approximation analysis which was carried out before the equation was chosen. Starting from scratch, the choice of an approximately correct equation would have been very difficult.

JOINT *vs.* ADDITIVE ANALYSIS

When there are joint relationships, joint correlation has the advantage over additive correlation in that it shows the facts more accurately. Additive correlation assumes that the relationships between two variables are independent of the size of a third variable. As observed in the examples in this chapter, this assumption is not always true.[9] Joint analysis is more flexible than additive. Joint correlation permits the effect of one variable to change with changes in other variables.

This greater flexibility in joint correlations is also a disadvantage. With the same number of observations, rho and the relationship are less reliable in joint than in additive correlation. This is especially true for the results of approximation analysis. Another disadvantage inherent in joint correlation is the difficulty in *showing* the relationships. Additive relationships can be shown more effectively in either graphic or tabular form than joint relationships.

---

[8] Pages 250 to 257.

[9] In fact, the assumption is almost never true, but in many problems the joint relationships are not sufficiently distinct to hamper seriously the use of additive correlation.

USES

Joint correlation methods have not been used so extensively as additive correlation methods. Even where relationships were definitely joint, the use of joint correlation has been limited by the complexity of both the problem and the method. When joint relationships have existed, there has been a tendency on the part of the research worker to employ tabular analysis rather than correlation analysis.

A few students have successfully studied problems with joint correlation methods. In studying the per capita consumption of milk, Waugh[10] analyzed linear joint relationship with the least-squares method. He found that size of family income was jointly related to consumption of milk in Boston. The author stated the relationship in the form of a linear joint regression equation as follows:

$$\text{Consumption} = 0.703 - 0.1285\ (\text{size family}) + (0.03408\ \text{size family} - 0.0614)\ \text{income}$$

The joint factor in the equation was (0.03408 size family)(income). Apparently, with high incomes, per capita consumption of milk did not vary with size of family. With small incomes, per capita consumption decreased as the families became larger.

Raeburn[11] used linear and curvilinear joint analysis by the approximation method in his study of quality and the price of apples. He found that defects and size were jointly related to price. With no defects, medium-sized apples brought more than large ones. When many defects were present, retailers paid 21 cents a bushel more for large than for medium-sized apples. Defects of the same size when cut out of large apples resulted in a smaller proportionate waste than when cut out of small apples. Although the medium-sized apple was preferred when sound, the greater proportionate waste penalized this size more than large sizes, when the quality was poor.

Underwood[12] used approximation analysis in studying the joint relation of acreage, yield, and price to returns from raising flue-cured tobacco. High yields increased returns per hour more when prices and acreage were high than when low. Increases in acreage increased returns more when yields and prices were high than when low. The joint relationship of the three variables to returns was shown as a series of three-dimensional graphs. The fourth dimension was represented by differences between the individual three-dimensional graphs.

---

[10] Waugh, F. V., The Consumption of Milk and Dairy Products in Metropolitan Boston in December, 1930, p. 12, September 1931.

[11] Raeburn, J. R., Joint Correlation Applied to the Quality and Price of McIntosh Apples, Cornell University Agricultural Experiment Station, Memoir 220, p. 20, March 1939.

[12] Underwood, F. L., Flue-Cured Tobacco Farm Management, Virginia Agricultural Experiment Station, Technical Bulletin 64, p. 98, January 1939.

MULTIPLE CORRELATION PROCEDURES

The student who undertakes a problem in multiple correlation should have constantly in mind the nature of the relationships and the methods to be employed. The nature of the relationship concerns (*a*) additive and joint effects, and (*b*) linearity and curvilinearity. The method of approach involves (*a*) least-squares and (*b*) approximation methods. There are eight different combinations of methods and types of relationships.

The procedure to be followed depends on the nature of the relationships. In general, the best methods to pursue might be summarized as follows:

Relationship		Usual Method	Result[13]	Chapter	Pages
Additive	Linear	Least-squares	Coefficient of multiple correlation, $R_{1.23}$ . . .	10	168 to 176
Additive	Curvilinear	Approximation, Ezekiel or short-cut	Index of multiple correlation, $\rho_{1.23}$ . . . (approximated)	13	217 to 239
Joint	Linear	Least-squares	Index of linear joint correlation, $\rho_{1.23}$ . . . (linear joint)	14	246 to 250
Joint	Curvilinear	Approximation, from contours	Index of curvilinear joint correlation, $\rho_{1.23(\text{curvilinear joint, contours})}$	14	252 to 257

Since there are so many different procedures that could be followed in any one problem, the student should make certain of the type of relationship and the most suitable method before becoming involved in detailed calculations.

[13] When the problem is completed, the measures of relationship should be clearly labeled in subscripts or footnotes, so that the reader will know what methods were used.

CHAPTER 15

TABULATION *vs.* CORRELATION ANALYSIS

During the past quarter of a century, there has been a controversy among agricultural statisticians concerning the relative merits of the tabulation and correlation methods of analyzing relationships. Most textbooks ignore this issue. In fact, most do not describe both methods of analysis. The emphasis is on correlation; tabulation is rarely described as a method of analyzing relationships. This is possibly due to the fact that the tabulation methods are considered too simple a tool for the "advanced" statistician and that the correlation method is much more difficult and therefore requires much explanation. In this book, six chapters are devoted to correlation and one chapter to tabulation analysis. The proportion of this and other textbooks devoted to the two methods of approach gives no clue as to their relative merits. An attempt is made in this chapter to compare results from the two methods and to summarize their advantages and disadvantages. The approach is first to analyze simple relationships and then to proceed step by step to more complicated problems of multiple relationship.

SIMPLE LINEAR RELATIONSHIPS

In studying a simple relationship, the analyst attempts to answer certain important questions approximately in the order of their importance.

1. Does a relationship exist?
2. Is the relationship positive or negative?
3. Is the relationship linear or curvilinear?
4. What is the rate of change in the dependent variable per unit change in the independent variable?
5. How closely are the two factors related?

TABULAR ANALYSIS

The records for 907 farms were classified according to an index of labor efficiency. The incomes were summed and averaged, and the averages were arranged in an orderly manner to facilitate comparison (table 1, left).

TABLE 1.—RESULTS OF TABULATION AND CORRELATION ANALYSIS OF A SIMPLE LINEAR RELATIONSHIP

RELATION OF LABOR EFFICIENCY TO INCOME ON 907 NEW YORK FARMS, 1927

Tabulation Analysis		*Correlation Analysis*
Index of efficiency X_4	Income X_1	Coefficient of correlation $r_{14} = +0.44$ Coefficient of determination $r^2_{14} = 0.19$ Regression equation $X_1 = +\$4.735X_4 - \$601$ (rising straight line)
Less than 100	\$ − 259	
100–199	+ 43	
200–299	+ 592	
300–399	+1,066	
400 and more	+1,400	

1. A relationship did exist between efficiency and incomes. This is shown by the fact that incomes changed more or less consistently with efficiency.

2. The relationship is positive; that is, incomes increased with increasing efficiency. This is shown by the fact that, for the least efficient farms, incomes were −\$259; and those for the most efficient, +\$1,400 (table 1, left).

3. The relationship is approximately linear; that is, each increase in efficiency resulted in about the same increase in incomes regardless of whether efficiency was high or low. This was shown by comparing the differences between average incomes for each successive group of farms. The average income for farms with lowest efficiency was −\$259; and for the next lowest group, +\$43 (table 1, left). The difference was +\$302. The other differences between other successive incomes were calculated in the same way. The four successive differences were as follows:

\$ +302
+549
+474
+334

For a relationship to be exactly linear, these differences should all be the same. They are not exactly the same, but there is no tendency for them to become consistently larger or consistently smaller. Therefore, the fluctuations among these differences are probably due to chance. For all practical purposes, the differences are about the same size, and the relationship may be assumed to be linear.

4. For each unit change in efficiency, incomes rose \$4.09. From the lowest to the highest efficiency groups, average incomes increased \$1,659, and the index of efficiency increased 406 units (481 − 75 = 406). The rate of increase[1] was \$1,659 divided by 406, or \$4.09.

5. It is not possible to state how closely the two factors are related.

CORRELATION ANALYSIS

The sums of squares and products for efficiency and income were obtained for the 907 farms. With the product-moment method of correlation without deviations,[2] the coefficients of correlation, determination, and regression and the regression equation were obtained. The results of correlation analysis, as usually presented, are given in table 1, right.

1. Some relationship did exist between efficiency and income. This is shown by the correlation coefficient $r_{14} = 0.44$.

2. The relationship was positive. This is indicated by the positive sign of the correlation coefficient, +0.44.

3. It is not possible to state whether the relationship is linear.

4. For each unit change in efficiency, incomes rose \$4.74. This is shown by the coefficient of regression in the equation $X_1 = +\$4.735X_4 - \$601$.

5. The two factors are not closely related. This is shown by the coefficient of determination, 0.19, which indicates that 19 per cent of the squared variability in income can be ascribed to differences in efficiency.

COMPARISON

1. Both methods show the existence of a relationship. In both cases, the decision as to whether there is a relationship depends on judgment. One decision is based on the trend in the averages; and the other, on the size of the coefficient. It often takes judgment to decide whether

---

[1] From the next lowest to the next highest efficiency groups, average income increased \$1,023 (1,066 − 43 = 1,023). The corresponding difference in efficiency was 185 units (331 − 146 = 185). The average rate of increase in income per unit of efficiency was \$5.53. This rate is probably more accurate than that based on the highest and lowest groups, because there is likely to be a larger number of farms in such classes than in the lowest and highest classes. The lowest and highest classes contained 69 and 37 farms, respectively, whereas the next lowest and next highest groups contained 392 and 101 farms, respectively. When the two lowest and two highest groups were weighted according to the number of farms, the rate of increase in income was \$4.93 per unit of efficiency.

[2] Page 153.

the trend in the averages is sufficiently consistent to indicate unquestionably the presence of a relationship. Likewise, it takes judgment to decide whether the correlation coefficient is large enough to indicate unquestionably the presence of a relationship.[3]

2. Both methods indicate that the relationship was positive.

3. The tabular method indicated that the relation was approximately linear. From the usual results of correlation analysis, it is not possible to reach any conclusion on this point. The regression equation indicates that the relationship is linear. However, this is due to the method, and not to the facts in the case. When the relationship appears linear with the tabular analysis, it *is linear*. When the relationship appears linear with correlation analysis, it *may or may not be linear.*

4. Both methods give the rate of change. The tabular method indicated that, for each unit change in efficiency, income increased \$4.09; and the correlation method, \$4.74. The rate of change by the correlation method is the more accurate. It is a weighted rate based on all the observations. By the common procedure of using only the highest and lowest classes, the rate of change by the tabular method is not likely to be very accurate. Accuracy can be increased by using a more refined procedure.[4]

5. Tabular analysis gives no indication of how closely the two factors are related. The correlation method gives a precise answer to this question.

The two methods of analyzing relationships may also be compared on the basis of (*a*) the amount of time required to obtain the results, (*b*) ease of presentation of results by the author, and (*c*) ease of interpretation by the layman. Tabular methods require much less time than correlation methods. Correlation results can be presented in fewer figures and less space than tabular results. From the standpoint of the layman, the tabular analysis has an overwhelming advantage. Coefficients of correlation and determination have little meaning for the layman. These coefficients are in abstract terms; whereas the results of tabular analysis are always in terms of dollars, cows, people, and the like. Mathematical equations also confound the layman. The regression equation $X_1 = +\$4.735X_4 - \$601$ is no exception.

However, this equation can be simplified by solving the equation for various values of X_4, as follows:

---

[3] These judgments can be tested statistically, pages 405 to 408. However, the usual practice is not to test them.

[4] Footnote 1, page 266.

EFFICIENCY,[5] X_4	INCOME, X_1
50	− 364
150	+ 109
250	+ 583
350	+1,056
450	+1,530

This shows that, with an increase of 100 in the index of efficiency, incomes rose $474. From the standpoint of the statistician, there is no difference between this table and the regression equation. From the standpoint of millions, the table is informative, while the equation is a riddle.

TABLE 2.—RESULTS OF TABULATION AND CORRELATION ANALYSIS OF A SIMPLE CURVILINEAR RELATIONSHIP

RELATION OF CROP YIELDS TO INCOMES ON 907 NEW YORK FARMS, 1927

Tabulation Analysis		*Correlation Analysis*
Index of crop yields X_3	Income X_1	Index of correlation $\rho_{(\text{freehand})} = 0.23$
Less than 60	$145	Coefficient of determination $\rho^2 = 0.05$
60– 79	171	
80– 99	251	
100–119	401	Freehand curve rising at an increasing rate.
120 or more	864	

SIMPLE CURVILINEAR RELATIONSHIPS

In analyzing simple curvilinear relationships, the analyst attempts to answer the following questions:

---

[5] If the group averages for efficiency were substituted in the equation, the results would be as follows:

EFFICIENCY X_4	INCOME, X_1 *estimated from regression*	INCOME, X_1 *calculated averages (table 1)*
75	$− 246	$− 259
146	+ 90	+ 43
238	+ 526	+ 592
331	+ 966	+1,066
481	+1,677	+1,400
Average	+ 375	+ 375

In some cases, the incomes estimated from the regression equation are higher than the actual averages; and in some cases, lower. The estimated averages typify the relationship shown by the actual averages under the assumption that the relationship is exactly linear. The weighted averages of the two series are, of course, the same.

1. Does a relationship exist?
2. Is the relationship positive or negative?
3. Is the relationship curvilinear?
4. What is the nature of the curvilinearity?
5. What are the rates of change in the dependent variable for different levels of the independent variables?
6. How closely are the two factors related?

TABULAR ANALYSIS

The records for 907 farms were classified according to crop yields, and the incomes were summed, averaged, and arranged in tabular form (table 2, left).

1. A relationship did exist between yields and income.

2. The relationship is positive.

3. The relationship is curvilinear. Incomes do not increase at a uniform rate as crop yields improve.

4. Incomes increase at an increasing rate as crop yields improve. This may be found by comparing the differences in successive income groups: +26, +80, +150, +463. The changes between successive groups increased rapidly as crop yields became better. The nature of this curvilinearity may also be found by plotting the data (figure 1).

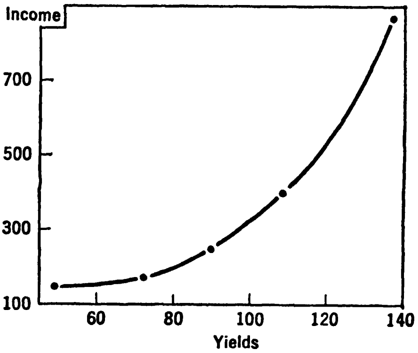


FIGURE 1.—RELATION OF YIELDS TO INCOME

The relationship is curvilinear. As crop yields improve, incomes rise at an increasing rate.

5. The rates of change in income at any given level of yields may be easily calculated. From the poorest crop yields to the next poorest yields, incomes increased $26. The index of yield increased 23 points (72 − 49 = 23). The rate of increase in income was $1.13 per point (26 ÷ 23 = 1.13). From the next best to the best yields, incomes increased $463, and the index of yields increased 28 points (137 − 109 = 28). Increases in good yields raised incomes $16.54 per point (463 ÷ 28 = 16.54).

6. It is not possible to state how closely the two variables are related.

CORRELATION ANALYSIS

The index of correlation was calculated by the freehand method.[6]

[6] Page 203.

1. There was a relationship between yields and income, $\rho = 0.23$ (table 2, right).

2. The relationship is positive. This is indicated by the shape of the curve.

3. The relationship is curvilinear.

4. The nature of the curvilinear relationship is that incomes increase at an increasing rate.

5. For low, average, and high yields, one point of change on the curve was accompanied by \$1, \$8, and \$23 increases in incomes.[7]

6. The two factors are not closely related. The coefficient of determination, $\rho^2 = 0.05$, indicates that only 5 per cent of the squared variability in income was explainable by differences in yields.

MULTIPLE RELATIONSHIPS, THREE VARIABLES

In analyzing problems involving two or more independent variables, tabular analysis presents the results in the form of two-way, three-way, and higher-order tables; and the correlation method, in the form of multiple correlation coefficients, indexes of multiple correlation, and regression equations and curves.

The multiple relation between size of farm, crop yields, and income was used for the comparison of the two methods of analysis (table 3).

Tabular Analysis

A multiple relation existed between size of farm, X_2, crop yields, X_3, and the dependent variable, income, X_1. With yields held constant, incomes increased regardless of whether crop yields were poor or good (table 3, top). As yields improved with size of farm held constant, incomes increased regardless of whether farms were small or large.

The net relationship of size of farm to income was linear. As farms became larger, incomes increased at an approximately constant rate. The linearity of this net relationship was tested in the following manner:

1. The incomes on small farms averaged −\$110, +\$31, and +\$202 for poor, medium, and good yields, respectively (table 3, top). The simple average of these three group averages was +\$41. Likewise, the simple averages for medium and large farms were +\$299 and +\$857, respectively.

2. The sizes of small farms averaged 147.5, 155.9, and 150.5 units

---

[7] According to the curve, incomes rose \$1 as yields increased from 59 to 60; \$8, from 99 to 100; and \$23, from 139 to 140.

TABLE 3.—RESULTS OF TABULATION AND CORRELATION ANALYSIS OF A MULTIPLE RELATIONSHIP INVOLVING THREE VARIABLES

Relation of Size of Farm and Crop Yields to Incomes on 907 New York Farms, 1927

Tabulation analysis

Size of farm, X_2	Crop yields, X_3		
	Poor	Medium	Good
	Income, X_1	*Income*, X_1	*Income*, X_1
Small	\$ −110	\$ + 31	\$ + 202
Medium	+238	+158	+ 500
Large	+600	+711	+1,261

Correlation analysis

Coefficient of multiple correlation $R_{1.23} = 0.48$.
Coefficient of determination $R^2_{1.23} = 0.23$.
Regression equation $X_1 = +\$2.12X_2 + \$6.17X_3 - \$887$.
Index of multiple correlation, ρ, too difficult to work.

for poor, medium, and good yields, respectively.[8] The simple average of these three groups was 151.3. Likewise, the simple averages for medium and large farms were 275.1 and 530.2, respectively.

3. The differences in income and in size of farms from small to medium and medium to large farms may be calculated as follows:

Size of Farm	Simple Average Income	Difference in Income	Simple Average Size	Difference in Size
Small	\$ 41		151.3	
		\$258		123.8
Medium	299		275.1	
		\$558		255.1
Large	857		530.2	

---

[8] The averages for size of farms were as follows:

Size of Farm	Crop Yields *Poor*	*Medium*	*Large*	Simple Average
Small	147.5	155.9	150.5	151.3
Medium	277.0	273.1	275.1	275.1
Large	480.0	537.5	573.0	530.2

4. By dividing the differences in income by the corresponding differences in size, the rate of change in income may be calculated as follows:

CHANGE IN SIZE	DIFFERENCES IN *Income*	DIFFERENCES IN *Size*	RATE OF CHANGE IN INCOME PER UNIT OF SIZE
Small to medium	$258	123.8	$2.08
Medium to large	558	255.1	2.19
Small to large	816	378.9	2.15

Since the rates of change from small to medium and from medium to large were practically the same, the relationship may be said to be linear. The average rate of change from small to large would be $2.15 per unit in size.

The net relationship of yields to income was curvilinear. As yields improved, incomes increased at an increasing rate. Based on the simple averages of incomes and yields and their differences, the rates of change in income from poor to medium and from medium to good crop yields were obtained as follows:

YIELDS	SIMPLE AVERAGE INCOME[9]	SIMPLE AVERAGE YIELDS[10]	DIFFERENCE IN *Income*	DIFFERENCE IN *Yields*	RATE OF CHANGE IN INCOME PER UNIT OF CROP YIELD
Poor	$243	71.2			
			$ 57	28.1	$2.03
Medium	300	99.3			
			354	27.7	12.78
Good	654	127.0			
Poor to good			$411	55.8	$7.37

The difference in income between medium to good crop yields was more than six times as large as that from poor to medium yields. Since the rates of change were decidedly different, there is little question but that the relationship between yields and income was curvilinear. The first increase in yields raised incomes only $2 per point, whereas the second increase raised incomes over $12. In other words, as yields increased, incomes increased at an increasing rate.

---

[9] Simple averages of the columns of table 3.

[10] The average indexes of crop yields were as follows:

SIZE OF FARM	CROP YIELDS *Poor*	CROP YIELDS *Medium*	CROP YIELDS *Good*
Small	67.6	98.8	129.1
Medium	73.5	99.1	126.1
Large	72.6	100.0	125.8
Simple average	71.2	99.3	127.0

The increase in income for each unit increase in crop yields depends on the level of the crop yields. The average rates of change determined above were +\$2.03 and +\$12.78. In reality, there may be many other rates of change depending on whether yields were increasing from very poor to poor, poor to fair—or good to excellent. It is doubtful whether a single rate of change in income with yields is valid where the relationship is curvilinear. However, an estimate of this rate of change, \$7.37, could be obtained by dividing the difference in income by that in yields, as yields improved from poor to good.

CORRELATION ANALYSIS

A multiple relationship existed between size, yields, and incomes, $R_{1.23} = 0.48$ (table 3, bottom). The linear net regression equation indicates that, on the average, incomes increased with size and with yields. However, it is not known whether incomes always increased with size for all levels of yields. Neither is it known whether incomes always increased with yields for all sizes of farms. In other words, it is not known whether these two relationships are linear or curvilinear.

The net linear rates of change were \$2.12 per unit change in size, and \$6.17 per unit change in yields. The validity of these two individual rates of change is dependent on whether the relationships are linear.

The coefficient of determination, $R^2_{1.23} = 0.23$, based on linear multiple analysis, measures the proportion of the squared variability in income explained by differences in size and yields.

An index of multiple curvilinear correlation from mathematically determined or freehand curves could be determined. However, with 907 farms, it is obvious that the amount of work involved in its calculation is prohibitive.

COMPARISON

Both methods indicate:

1. The existence of a multiple relationship.
2. Positive relationships between both independent variables and income.
3. The following approximately similar average rates of change:

METHOD	EFFECT OF SIZE	EFFECT OF YIELDS
Tabulation	\$2.15	\$7.37
Correlation	2.12	6.17

The tabular analysis indicated the following facts not revealed by correlation analysis:

1. Whether the relationships were linear or curvilinear.
2. The nature of the curvilinear relationship between yields and income.
3. The approximate rates of change for this curvilinear relationship at different levels of yields.

The correlation analysis indicated the following fact not revealed by tabular analysis:

1. The exact amount of the relationship.

INTERSERIAL RELATIONSHIPS

One of the troublesome problems of analyzing relationships arises when there are interrelationships among independent variables. Such interrelationships are called interserial relationships. There was a moderate amount of interserial relationship between size and yields of New York farms.

Tabular Analysis

This interrelationship cannot be detected in table 3 because only the average incomes are given. The assumption was made that large farms were always the same size regardless of yields, and good yields were equally good regardless of size. Consequently, the numerical measures of size and yields were not given. An interrelationship between size and yields can be seen in either a two-way or a one-way table, provided that averages for the independent variables are given.

For the two-way classification, the detailed information was given in table 4, top. From these data, an interrelationship might be suspected for two reasons:

1. Numbers of farms in the subgroups indicate that a larger proportion of yields were poor when farms were small than when large. These proportions would be

$$\text{Small farms: } \frac{139}{139 + 85 + 84} = 0.45 \qquad \text{Large farms: } \frac{101}{101 + 102 + 89} = 0.35$$

This indicates that yields are poorer on small than on large farms.

2. The average units for size indicate that, among all large farms, those with poor yields were smaller than those with good yields (compare 480.0 with 573.0).

Interrelationships are probably most easily detected in one-way tables. The average sizes and yields for one-way classifications by size and yield are given in table 4, bottom. Yields were higher on large than on small farms (compare 98.4 and 93.0). Likewise, farms with good yields were larger than those with poor yields (compare 334.3 with 284.1).

TABLE 4.—DETAILED INFORMATION IN TWO-WAY AND ONE-WAY TABLES NECESSARY TO DETECT INTERSERIAL RELATIONSHIPS

Relation of Size and Yields to Income on 907 New York Farms, 1927

Two-way table	Size of farm X_2	Crop yields X_3	Number of farms	Average units for		Income X_1
				Size X_2	Yields X_3	
	Small	poor	139	147.5	67.6	\$ − 110
	”	medium	85	155.9	98.8	+ 31
	”	good	84	150.5	129.1	+ 202
	Medium	poor	112	277.0	73.5	+ 238
	”	medium	97	273.1	99.1	+ 158
	”	good	98	275.1	126.1	+ 500
	Large	poor	101	480.0	72.6	+ 600
	”	medium	102	537.5	100.0	+ 711
	”	good	89	573.0	125.8	+1,261
	Small	—	308	150.6	93.0	—
One-way tables	Medium	—	307	275.1	98.4	—
	Large	—	292	528.4	98.4	—
	—	poor	352	284.1	71.0	—
	—	medium	284	333.0	99.3	—
	—	good	271	334.3	126.9	—

When interserial association is present, independent variables are not held entirely constant in a two-way or higher-order table. For example, when farms were large, an improvement in yields from poor to good raised income from \$600 to \$1,261 (table 4, top). However, this increase was due not only to an increase in yields but also partly to an increase in size. Size was not held constant; in fact, it increased from 480.0 to 573.0 (table 4, top).

The rates of change in income by tabular analysis, which were \$2.15 per unit of size and \$7.37 per point of crop index, were calculated with the assumption that in a two-way table the effects of independent variables are held constant. When interrelationships are present, this is not entirely true. In such cases, net rates of change may be more accurately calculated by a different procedure, as follows:

1. From simple averages, the differences in size, yields, and income resulting from increasing size from small to large and yields from poor

to good were calculated. These averages and differences were calculated from table 4, top, and set down in an orderly manner as follows:

INDEPENDENT VARIABLES	SIMPLE AVERAGES AND DIFFERENCES		
	Size	*Yields*	*Income*
Size of farm			
Small	151.3	98.5	$+ 41
Large	530.2	99.5	+857
Difference	+378.9	+ 1.0	$+816
Crop yields			
Poor	301.5	71.2	+243
Good	332.9	127.0	+654
Difference	+ 31.4	+ 55.8	+411

2. Each group of differences was set in equation form as follows:

$$\begin{pmatrix}\text{Difference in}\\ \text{size}\\ X_2\end{pmatrix}\begin{pmatrix}\text{Net rate of change in}\\ \text{income per unit}\\ \text{of size}\end{pmatrix}+\begin{pmatrix}\text{Difference in}\\ \text{yields}\\ X_3\end{pmatrix}\begin{pmatrix}\text{Net rate of change in}\\ \text{income per point}\\ \text{of yields}\end{pmatrix}=\begin{matrix}\text{Difference in}\\ \text{income}\\ X_1\end{matrix}$$

$$+378.9b_2 \qquad + 1.0b_3 \qquad = +816$$

$$+ 31.4b_2 \qquad +55.8b_3 \qquad = +411$$

3. These equations were solved simultaneously for the values of b_2 and b_3, the net rates of change.[11]

The net rate of change due to size was +$2.14; and that due to yields, +$6.16.

CORRELATION ANALYSIS

The procedure in correlation analysis is the same regardless of whether interserial relationships are present. The regression equation given in table 3, bottom, shows the net effects of size and yields on income.

COMPARISON

When interrelationships are present, the subgrouping of the tabular method fails to hold constant the effects of independent variables. Further analysis of averages is then necessary to determine the real net effects. On the other hand, correlation analysis indicates the net effect equally well whether or not these interrelationships are present.

The rates of change in income with size were remarkably similar by all methods (table 5). The first rate of change with yield by the tabulation method was $7.37, compared with $6.17 by correlation. However, the second rate of change, $6.16, determined from the averages with the aid of simultaneous equations, was practically the same as that from correlation.

---

[11] A suitable method for solving these equations is given in table 2, page 147

TABLE 5.—NET RATES OF CHANGE BY TABULATION AND CORRELATION METHODS

RELATION OF SIZE AND YIELD TO INCOME ON 907 NEW YORK FARMS, 1927

Method	Effect of size	Effect of yield
Tabulation		
Simple averages only	$2.15	$7.37
Simultaneous equations	2.14	6.16
Correlation	2.12	6.17

MULTIPLE RELATIONSHIPS, FOUR VARIABLES

TABULAR ANALYSIS

The 907 farms were classified and subclassified by size, X_2, yields, X_3, and efficiency, X_4. With 907 farms and 27 subgroups, some persons might assume that each group would contain about 30 farms. The exact distribution is shown in table 6. It is obvious that averages for six of the subgroups with 0, 0, 1, 6, 6, and 6 farms, respectively, would either be lacking or be based on insufficient data. This greatly limits the use-

TABLE 6.—DISTRIBUTION OF THE NUMBER OF FARMS CLASSIFIED AND SUBCLASSIFIED ACCORDING TO THREE INDEPENDENT VARIABLES WITH THREE GROUPS EACH

SIZE, CROP YIELDS, AND LABOR EFFICIENCY, 907 NEW YORK FARMS, 1927

Size of farm X_2	Crop yields X_3	Labor efficiency, X_4		
		Low	Medium	High
		Number of farms	*Number of farms*	*Number of farms*
Small	poor	99	40	0
"	medium	62	23	0
"	good	60	23	1
Medium	poor	13	54	45
"	medium	19	49	29
"	good	27	41	30
Large	poor	6	19	76
"	medium	6	31	65
"	good	6	36	47

fulness of the three-way table of averages that might be constructed. For example, the effect of high over medium efficiency on small farms could not be observed.

The unequal distribution of farms in table 6 was due to a marked interrelationship between size of farm and labor efficiency. Most small farms had low efficiency; and most large farms, high efficiency. The difficulty of unequal distribution arises whenever such interrelationships are present. This difficulty limits the applicability of tabular analysis to multiple relationships involving three or more independent variables. The difficulty may usually be overcome in one of two ways:

1. Increasing the total number of observations.
2. Decreasing the number of classes for each independent variable, that is, increasing the size of the subgroups.

In the income problem, the second method was used. For each of the three independent variables, the farms were divided into two approximately equal groups as follows:

SIZE OF FARM, X_2	CROP YIELDS, X_3	EFFICIENCY, X_4
Small	poor	low
Large	good	high

This reduced the number of subgroups from 27 to 8, and increased the number of farms[12] in the smallest group from 0 to 45.

The average incomes were obtained for the eight subgroups (table 7).

With size and crop yields held constant, incomes were related to efficiency. The net effect of efficiency may be observed by comparing the corresponding incomes in the two columns of table 7. On small farms with poor yields, incomes rose from −$119 to $384 with increasing efficiency. Similarly, when size and crop yields were held constant at other levels, incomes also rose with increasing efficiency.

With size and efficiency held constant, incomes were related to crop yields. The net effect of yields may be observed by comparing the first with the second and the third with the fourth rows of table 7. On small farms with low efficiency, incomes increased from −$119 to +$101 as crop yields improved.

---

[12] The numbers of farms in subgroups, which were still unequally divided, were as follows:

SIZE FARM X_2	CROP YIELDS X_3	LABOR EFFICIENCY, X_4	
		Low	*High*
Small	poor	181	49
”	good	166	46
Large	poor	45	173
”	good	68	179

TABLE 7.—RESULTS OF TABULATION AND CORRELATION ANALYSIS OF A MULTIPLE RELATIONSHIP INVOLVING FOUR VARIABLES

RELATION OF SIZE OF FARM, CROP YIELDS, AND LABOR EFFICIENCY TO INCOME ON 907 NEW YORK FARMS, 1927

*Tabulation analysis**

Size of farm X_2	Crop yields X_3	Labor efficiency, X_4	
		Low	High
		Income, X_1	*Income,* X_1
Small	poor	\$ −119	\$ + 384
"	good	+101	+ 361
Large	poor	−271	+ 592
"	good	+232	+1,139

Correlation analysis

Coefficient of multiple correlation, $R_{1.234} = 0.53$.
Coefficient of determination, $R^2_{1.234} = 0.28$.
Regression equation, $X_1 = +\$1.32X_2 + \$6.85X_3 + \$2.95X_4 - \$1{,}309$.
Partial coefficients, $r_{12.34} = 0.25$; $r_{13.24} = 0.20$; $r_{14.23} = 0.25$.
Beta coefficients, $\beta_{12.34} = 0.27$; $\beta_{13.24} = 0.18$; $\beta_{14.23} = 0.27$.

* Calculated from data given in Appendix D.

With crop yields and efficiency held constant, size was related to income. The net effect of size may be observed by comparing the first with the third and the second with the fourth rows of table 7. When yields were poor and efficiency low, large farms lost more, −\$271, than small farms, −\$119. However, when yields were good and efficiency high, large farms returned more income, +\$1,139, than small farms, +\$361. With other combinations of yields and efficiency, incomes rose with size of farm.

Some of the independent variables were jointly related to income with other independent variables. For example, the effect of size varied as follows:

$$\begin{array}{rcrcr}
\begin{pmatrix}\text{Large}\\ \text{farms}\end{pmatrix} & - & \begin{pmatrix}\text{Small}\\ \text{farms}\end{pmatrix} & = & \begin{pmatrix}\text{Effect}\\ \text{of size}\end{pmatrix}\\
\$-\ 271 & - & (\$ -119) & = & \$ -152\\
232 & - & 101 & = & +131\\
592 & - & 384 & = & +208\\
1{,}139 & - & 361 & = & +778
\end{array}$$

The effect of size, which varied from −\$152 to +\$778, depended on the combination of yields and efficiency (table 8). In other words, size was related to income jointly with some other factors.

TABLE 8.—EFFECTS OF SIZE OF FARM ON INCOME WITH THE EFFECTS OF OTHER VARIABLES HELD CONSTANT

DIFFERENCES IN GROUP INCOMES*

Efficiency X_4	Crop yields X_3	Differences in income, X_1, due to size of farm, X_2
Low	poor	\$ −152
"	good	+131
High	poor	+208
"	good	+778
Average or additive effects of size		+241

* Calculated from averages in table 7, page 279, or data in Appendix D.

The net effect of yields, which varied from −\$23 to +\$547, was probably jointly related with some other factor (table 9, left). The net effect of efficiency, which varied from +\$260 to +\$907, also indicated the presence of joint relationship.

The average net effects of size, +\$241, may be obtained by averaging the four differences in income due to size (table 8). This average effect is merely the difference between incomes on the half of the farms that were smaller and the half of the farms that were larger, with the effects of the other variables held constant. The average effect may also be called the "additive" effect of size on income. Stated another way, the change in size "added on the average" \$241 to incomes.

Similarly, the average net effect or "additive" effect of good over poor yields was +\$312; and of high over low efficiency, +\$633 (table 9). The size of the average effects, +\$241, +\$312, and +\$633, indicates the relative importance of the three factors in determining income. However, these average effects are chiefly valuable for further analysis.

The net rate of change in income with size of farm may be obtained by dividing the average effect of large over small size, +\$241, by the corresponding amount of increase in size, 217 units (table 10). The net rate of change was \$1.11 ($241 \div 217.1 = 1.11$). That is, with efficiency and crop yields held constant, incomes rose \$1.11 for each

TABLE 9.—EFFECTS OF (*a*) YIELDS AND (*b*) EFFICIENCY WHEN EFFECTS OF OTHER VARIABLES ARE HELD CONSTANT*

(*a*) Effects of *yields* with X_2 and X_4 held constant			(*b*) Effects of *efficiency* with X_2 and X_3 held constant		
Size of farm X_2	Efficiency X_4	Differences in income, X_1, due to yields	Size of farm, X_2	Crop yields X_3	Differences in income, X_1, due to efficiency
Small	low	$ +220	Small	poor	$ +503
"	high	− 23	"	good	+260
Large	low	+503	Large	poor	+863
"	high	+547	"	good	+907
Average or additive effects of yields		+312	Average or additive effects of efficiency		+633

* Calculated from averages in table 7, page 279, or data in Appendix D.

unit increase in size. The net rate of change in income with yields was $7.86 (312 ÷ 39.7 = 7.86); and with efficiency, $5.43 (633 ÷ 116.5 = 5.43). These rates of change might be arranged in an equation as follows:

$$X_1 = 1.11X_2 + 7.86X_3 + 5.43X_4 + \text{Constant}$$

These rates of change and "average effects" are always interesting. They are especially useful when the relationships are additive. However, when relationships are joint, average rates or average effects are inadequate.

TABLE 10.—SUPPLEMENTARY INFORMATION NECESSARY TO OBTAIN RATES OF CHANGE*

Efficiency X_4	Crop yields X_3	Differences in units from small to large farms	Size X_2	Efficiency X_4	Differences in units from poor to good yields	Size X_2	Yields X_3	Differences in units from low to high efficiency
Low	poor	194.2	Small	low	45.1	Small	poor	95.9
"	good	196.2	"	high	40.2	"	good	102.7
High	poor	205.6	Large	low	34.9	Large	poor	140.0
"	good	272.4	"	high	38.4	"	good	127.3
Average difference		217.1	Average difference		39.7	Average difference		116.5

* Calculated from data given in Appendix D.

The average effect of size was +\$241 (table 8). However, the effect of size with different combinations of yields and efficiency varied from −\$152 to +\$778. The four individual differences are more descriptive and probably more useful in describing the relation of size to income than is the average effect +\$241.

Likewise, rates of change based on the four individual differences are probably more useful than the rate based on the average, \$1.11. When efficiency was low and yields poor, the effect of size, −\$152, accompanied an increase of 194.2 units in size. The average rate of change was −\$0.78 $(-152 \div 194.2 = -0.78)$. Likewise, when efficiency was high and yields were good, the rate of change was +\$2.86 $(+778 \div 272.4 = +2.86)$. Since these two rates of change were very different, the average rate, +\$1.11, has little significance. The average rates of change with efficiency and yields are likewise inadequate.

The tables give no clue to the closeness of the relationships or to whether the relationships are linear or curvilinear.

CORRELATION ANALYSIS

There is no question but that a multiple relationship existed, $R_{1.234} = 0.53$ (table 7, bottom). Furthermore, each independent variable was positively related to income. This is shown by the positive signs of the three regression coefficients. The rates of change were \$1.32 for size, \$6.85 for yields, and \$2.95 for efficiency.

The multiple correlation analysis gives a definite indication of the closeness of the relationship, $R^2_{1.234} = 0.28$. The partial correlation coefficients indicate that there was a relationship between income and each independent variable after the effects of the other independent variables had been considered. The beta coefficients indicate the relative importance of the size, yields, and efficiency in determining income. Size and efficiency were about equally important, and both were more important than yields.

With correlation, the assumption is made that the relationships are linear and it is not revealed whether they are curvilinear. With 907 farms, curvilinear methods of correlation would involve an enormous amount of work. With correlation, it is also usually assumed that the effect of one independent variable has no relation to the effects of the other independent variables. All relationships are assumed to be additive, and it is not learned whether they are joint.

COMPARISON

Both methods indicate:

1. The existence of relationships between each independent variable and income.

2. Whether these average relationships were positive or negative.
3. The relative importance of each independent variable in determining income.
4. The average rates of change in income with changes in each independent variable.

Tabular analysis indicated the following fact not revealed by correlation analysis: whether the effects of each independent variable on income varied with different combinations of the other two independent variables; in other words, whether each factor was related to income jointly with some other factor.[13]

Correlation analysis indicated the following facts not revealed by tabular analysis:

1. The exact amount of the multiple relationship between the independent variables and the income.
2. The amount of relationship between each independent variable and income, in addition to the effects of the other independent variables.

Neither correlation nor tabulation analysis revealed whether the relationships were linear or curvilinear. For tabular analysis, there were not enough observations; and for correlation, there were too many.

ADDITIVE AND JOINT RELATIONSHIPS

Joint relationships are an important problem that has not been adequately treated by either method of analyzing relationships. The following discussion deals with the further analysis of joint relationships.

Tabular Analysis

Relationships shown by tabular analysis may be classified as either joint or additive. Additive relationships are merely average relationships. The additive effects of large over small farms were +\$241; of good over poor yields, +\$312; and of high over low efficiency, +\$633 (tables 8 and 9).

Joint relationships concern the variability in the effect of an independent variable on income with different combinations of the other independent variables. Stated another way, the problem of analyzing joint relationships is to measure whether and how the effect of one independent variable on the dependent changes with different values of the other independent variables.

---

[13] Correlation analysis holds independent variables constant at *one* level, *the average*. Tabular analysis holds the independent variables constant at *two* (or more) levels.

Joint Effect of Size and Efficiency on Income

Since size of farm, X_2, was probably most jointly related with other factors, the differences due to X_2 were chosen for detailed analysis (table 11). The first three columns in table 11 are identical to table 8. In the fourth column, the differences in income, X_1, due to size of farm, X_2, are compared for low and high labor efficiency, X_4. For the first two differences in the third column, labor efficiency was low; and for the last two, high. The conditions for the first first difference, −$152, were the same as for the third first difference, +$208, except for labor efficiency. Similarly, the second first difference, in the third column, +$131, was comparable to the fourth, +$778. The first difference, −$152, for low efficiency was subtracted from the comparable first difference, +$208, for high efficiency. The resulting difference in first differences, +$360, was called a second difference due to size of farm, X_2, and labor efficiency, X_4 (fourth column, table 11). The other second difference in income due to size and efficiency, +$647, was calculated similarly from the other two first differences (778 − 131 = 647).

TABLE 11.—ANALYSIS OF JOINT RELATIONSHIPS WITH SECOND DIFFERENCES

SECOND DIFFERENCES MEASURING THE JOINT EFFECT OF SIZE AND EFFICIENCY ON INCOMES

Efficiency X_4	Crop yields X_3	First difference* in income, X_1, due to size of farm, X_2	Second differences due to size, X_2, and efficiency, X_4
Low	poor	$ −152	..... $ +360
"	good	+131	
High	poor	+208	.... +647
"	good	+778	
Additive effect of size		+241	
Joint effect of size and efficiency			+504

* Table 8.

The second differences measure the extent to which the effect of size of farm is different for high and low efficiency. The average second difference was +$504. Size of farm increased income $504 more when efficiency was high than when low. Another interpretation would be that efficiency increased income $504 more when farms were large than when they were small. Either interpretation would be correct. Regardless of its interpretation, the average second difference is a measure of joint relationship between size and efficiency.

Joint Effect of Size and Yields on Income

Second differences due to size, X_2, and yields, X_3, were calculated by subtracting the first differences due to yields when farms were small from the corresponding first differences due to yields when farms were large. These second differences averaged +$427 (table 12). When crop yields were good, size of farm increased incomes $427 *more* than when crops were poor. This indicates a joint relation of size and yields to income. The average second difference, +$427, is a measure of the amount of this joint relationship.

TABLE 12.—ANALYSIS OF JOINT RELATIONSHIPS
WITH SECOND DIFFERENCES

SECOND DIFFERENCES MEASURING THE JOINT EFFECTS OF YIELDS AND EFFICIENCY AND OF YIELDS AND SIZE ON INCOME

Size of farm X_2	Efficiency X_4	First differences* in income, X_1, due to yields, X_3	Second differences in income, X_1, due to	
			Yields, X_3, and efficiency, X_4	Yields, X_3, and size, X_2
Small	low	$ +220	$ −243	$ +283
"	high	− 23		+570
Large	low	+503	+ 44	
"	high	+547		
Additive effects of yields		+312		
Joint effects of yields and efficiency			−100	
Joint effects of yields and size				+427

* Table 9, left.

Joint Effect of Yields and Efficiency on Income

The joint effect of yields and efficiency on incomes was −$100 (table 12). In other words, when efficiency was high, the improvement from poor to good yields raised incomes $100 less than when efficiency was low. However, it is doubtful whether this second difference is large enough to be significant.

TABLE 13.—ANALYSIS OF JOINT RELATIONSHIPS WITH THIRD DIFFERENCE

THIRD DIFFERENCE MEASURING THE JOINT EFFECT OF SIZE, YIELDS, AND EFFICIENCY ON INCOME

Size of farm X_2	Efficiency X_4	First differences* in income, X_1, due to yields, X_3	Second differences in income, X_1, due to yields, X_3, and efficiency, X_4	Third difference in income, X_1, due to yields, X_3, efficiency, X_4, and size, X_2
Small	low	$ +220	....$ −243	....$ +287
"	high	− 23		
Large	low	+503	.... + 44	
"	high	+547		
Additive effects of yields		+312		
Joint effects of yields and efficiency			−100	
Joint effects of yields, efficiency, and size				+287

* Table 9, left.

Joint Effect of Size, Yields, and Efficiency on Income

Sometimes, there are joint relationships among three or more independent and the dependent variables. The joint effect of size, yields, and efficiency is measured by the third difference due to those three variables. The third difference in income can be calculated from the second differences in the same way that the second were obtained from the first differences.[14] Third differences due to size, yields, and efficiency were calculated from second differences due to X_3 and X_4 (table 13). The second difference for small farms, −\$243, was subtracted from the second difference for large farms, +\$44. The resulting third difference, +\$287, measures the joint effect of size, yields, and efficiency.

---

[14] Each of the three average second differences shown in tables 11 and 12 could have been calculated in another way. For example, in table 11, the differences in income due to size of farm, X_2, were calculated first; and then the differences in these differences due to efficiency, X_4, were obtained. These second differences could have been obtained by considering size and efficiency in the reverse order, beginning with first differences due to efficiency in table 9, right. The second difference would be exactly the same, regardless of whether size or efficiency was considered first.

Likewise, the third difference, +$287, obtained in table 13 could have been calculated six different ways, depending on the order in which three independent variables were considered.

The third difference, +\$287, measures the effect of size on the effect of efficiency on the effect of yields on income. When farms were large, increased labor efficiency increased the effect of yields on income \$287 more than when farms were small.

This third difference, +\$287, also measures the effect of yields on the effect of size on the effect of efficiency on X_1. When yields were good, size of farm increased the effect of efficiency \$287 more than when yields were poor.

A third difference may be interpreted as any one of six different combinations of effects on effects on effects. All six interpretations would be equally correct and equally confusing. The choice of the interpretation depends on the student's interest. The third difference, +\$287, is a single number which expresses a complicated three-way joint relationship. Its simplicity is limited by the capacity of the human mind to conceive of such relationships.

Summary of Additive and Joint Relationships

All the possible additive and joint relationships in a three-way table have now been examined. Additive relationships were measured by average first differences; and joint relationships, by average second and third differences. A summary of these measures may give some idea as to which relationships were most important (table 14).

However, the first, second, and third differences are not directly comparable. The first difference, +\$241, due to size of farm is not directly comparable to the second difference, +\$427, due to size of farm and crop yields (table 14). Additive and joint effects may be compared directly by converting the differences into effects on average income. The additive effect of large over average-sized farms, +\$121 is one-half the effect of large over small farms, \$241. The joint effect of large over average-sized farms and good over average yields, \$107, is only one-fourth of the second difference, \$427. The first to the third differences were divided by 2, 4, and 8, respectively, to obtain the effects on average income. The inconsistency in the divisor was necessitated by the peculiarities of the successive differences method.

The relative importance of the additive relationship, size to income, and the joint relationships, size and yields to income, is in the proportion of \$121 to \$107, not \$241 to \$427.

Labor efficiency had the greatest additive effect on income, +\$317, followed by yields, +\$156, and size of farm, +\$121. The additive

TABLE 14.—SUMMARY OF ADDITIVE AND JOINT RELATIONSHIPS FROM THREE-WAY TABULAR ANALYSIS

RELATIONSHIPS OF SIZE OF FARM, CROP YIELDS, AND LABOR EFFICIENCY TO INCOME ON 907 NEW YORK FARMS, 1927

Independent variables	Average differences	Effect on average income*	Estimated degree of relationship†
Additive relationships	*First differences*		
Size of farm, X_2	$ +241	$ +121	Marked
Crop yields, X_3	+312	+156	Marked
Labor efficiency, X_4	+633	+317	Marked
Joint relationships, two-way	*Second differences*		
Size of farm and crop yields, X_2X_3	+427	+107	Marked
Size of farm and efficiency, X_2X_4	+504	+126	Marked
Crop yields and efficiency, X_3X_4	−100	− 25	Doubtful
Joint relationships, three-way	*Third difference*		
Size of farm, crop yields, and efficiency, $X_2X_3X_4$	+287	+ 36	Doubtful

* The first differences are twice the deviation from the average; second differences based on first differences were four times the deviation from the average; and the third differences, eight times.

† The amount of the difference due to random fluctuation was not studied at this point. It was arbitrarily decided that, when effects were $100 or more, the relationships were "marked"; when $50 to $100, "definite"; $20 to $50, "doubtful"; and less than $20, "none." A criterion of reliability of differences based on variability in income will be discussed on page 386.

effects of efficiency, yields, and size were great enough to be considered marked.

Among the two-way joint relationships, two were marked and the third was doubtful. The largest joint effect, +$126, indicated that either size of farm or labor efficiency increased income more when the other was large than when small. Likewise, size of farm and crop yields had a definite joint effect in addition to their additive effects. The three-way joint relationship was doubtful.

Apparently, the additive effects were somewhat more important than the joint effects.

Rates of Change

The foregoing analysis of additive and joint effects is simple and is satisfactory for most purposes. However, the effects of interserial relationships have not been taken into account with this method. In calculating additive and joint rates of change in income, more accurate results may be obtained by a different procedure as follows:

1. For each subgroup in table 7, income was expressed in terms of the three independent variables in an equation.

$$\begin{pmatrix}\text{Size of}\\\text{farm}\\X_2\end{pmatrix}\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\\text{size}\end{pmatrix}+\begin{pmatrix}\text{Crop}\\\text{yields}\\X_3\end{pmatrix}\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\\text{yields}\end{pmatrix}+\begin{pmatrix}\text{Effi-}\\\text{ciency}\\X_4\end{pmatrix}\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\\text{efficiency}\end{pmatrix}+(X_2X_3)\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\X_2\text{ and }X_3\end{pmatrix}$$
$$+(X_2X_4)\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\X_2\text{ and }X_4\end{pmatrix}+(X_3X_4)\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\X_3\text{ and }X_4\end{pmatrix}+(X_2X_3X_4)\begin{pmatrix}\text{Net rate}\\\text{of change}\\\text{due to}\\X_2, X_3,\text{ and }X_4\end{pmatrix}+\text{Constant}=\text{Income}$$

For small farms with poor crops and low efficiency, this equation[15] was

$$157.6b_2 + 73.1b_3 + 125.7b_4 + (157.6)(73.1)b_{23} + (157.6)(125.7)b_{24} + (73.1)(125.7)\,b_{34} + (157.6)(73.1)(125.7)\,b_{234} + C = -119$$

or, more simply,

$$157.6b_2 + 73.1b_3 + 125.7b_4 + 11{,}521b_{23} + 19{,}810b_{24} + 9{,}189b_{34} + 1{,}448{,}134b_{234} + C = -119$$

Seven similar equations were made for each of the other combinations of size, yields, and efficiency shown in Appendix D.

2. The eight equations were solved simultaneously for the seven unknown rates of change and the constant.[16] These values were

Additive effect of size, $b_2 = -5.480$
” ” ” yields, $b_3 = +16.643$
” ” ” efficiency, $b_4 = +18.990$

Joint effect of size and yields, $b_{23} = +0.02865$
” ” ” size and efficiency, $b_{24} = -0.01456$
” ” ” yields and efficiency, $b_{34} = -0.17221$
” ” ” size, yields, and efficiency, $b_{234} = +0.00026223$
Constant $C = -1{,}698$

These values established the following equation of relationship:

$$X_1 = -5.480X_2 + 16.643X_3 + 18.990X_4 + 0.02865X_2X_3 - 0.01456X_2X_4 - 0.17221X_3X_4 + 0.00026223X_2X_3X_4 - 1{,}698$$

In this equation, as in any equation showing joint relationships, the individual rates of change are practically meaningless. They take on meaning only when the levels at which independent variables are held constant are given. For example, in studying the effect of size, if the index of yields is assumed to be good, 120, and efficiency is assumed to be high, 350, the equation simplifies to

---

[15] The values of the independent variables needed for this equation are given in Appendix D, page 433.

[16] A suitable method for solving these equations is given on page 174.

$$X_1 = -5.480X_2 + 1{,}997 + 6{,}647 + 3.438X_2 - 5.096X_2 - 7{,}233 + 11.014X_2 - 1{,}698$$

$$X_1 = +3.88X_2 - 287$$

On the other hand, if yields are poor, 80, and efficiency is low, 150, the equation simplifies to

$$X_1 = -5.480X_2 + 1{,}331 + 2{,}849 + 2.292X_2 - 2.184X_2 - 2{,}067 + 3.147X_2 - 1{,}698$$

$$X_1 = -2.23X_2 + 415$$

The net effect of size, X_2, is different for different levels of yields and efficiency. When efficiency and yields are high, large farms return more income than small farms. When efficiency and yields are low, the reverse is true.

The various additive and joint relationships are most easily studied by converting the equation into terms of deviations from the averages of size, yields, and efficiency, as follows:

$$X_1 = -5.480(x_2 + AX_2) + \cdots + 0.02865(x_2 + AX_2)(x_3 + AX_3) + \cdots + 0.00026223(x_2 + AX_2)(x_3 + AX_3)(x_4 + AX_4) - 1{,}698$$

Collecting terms,

$$X_1 = -0.512x_2 + 6.856x_3 + 5.621x_4 + 0.08157x_2x_3 + 0.01086x_2x_4 - 0.09195x_3x_4 + 0.00026223x_2x_3x_4 + 222$$

This equation gives rates of change comparable to the additive and joint effects on average income given in table 14. These rates of change are those which hold for any independent variable or variables when the remaining independent variables are held constant at their averages. According to this equation, the average or additive rate of change in income with size, X_2, was $-\$0.51$ per unit. This indicates that, if yields and efficiency were average, an increase of 1 unit in size would reduce income by \$0.51. This change does not hold if yields or efficiency or both change simultaneously with size. For example, if efficiency, X_4, were held constant at its average, and size increased 50 units above average, and yields increased 10 points above average, income would increase \$84. The increase might be allocated to the additive and joint effects as follows:[17]

---

[17] The validity of subdividing the total effect into various additive and joint effects is questionable. The additive effect of size, $-\$25.60$, is that which would have held for a change of 50 units in size with yields and efficiency held constant. However, yields were not held constant. Many persons prefer to consider \$84 the joint effect of size and yields. With such a terminology, effects may be either joint or additive, but not partly both.

Additive effect of size, X_2	= \$ -25.60	$(50 \times -0.512 = -25.60)$
Additive effect of yields, X_3	= $+68.56$	$(10 \times +6.856 = +68.56)$
Joint effect of size and yields, X_2 and X_3	= $+40.79$	$(50 \times 10 \times +0.08157 = +40.79)$
Total	$+83.75$	

However, the total effect of size would not be −\$25.60, but this amount plus an indeterminate part of the joint effect, +\$40.79.

Equations of relationship are often useful in estimating incomes for different combinations of independent variables. Either equations in terms of actual values or those in terms of deviations can be used. However, those in terms of actual values are the more useful for this purpose, because deviations need not be calculated. The determination of the estimated income consists of substituting the given values of size, yields, and efficiency in the equation and simplifying and collecting the resulting terms. The estimated income for small farms (200 units), good yields (125 points), and average efficiency (200 units) would be calculated as follows:

$$\begin{aligned} X_1 &= -5.480(200) + 16.643(125) + 18.990(200) + 0.02865(200 \times 125) \\ &\quad - 0.01456(200 \times 200) - 0.17221(125 \times 200) + 0.00026223(200 \\ &\quad \times 125 \times 200) - 1{,}698 \\ &= -1{,}096 + 2{,}080 + 3{,}798 + 716 - 582 - 4{,}305 + 1{,}311 - 1{,}698 \\ X_1 &= +224 \end{aligned}$$

CORRELATION ANALYSIS

The usual method of analyzing multiple relationships by correlation is to calculate some "additive" measure of correlation, such as the multiple correlation coefficient, R, or such as the indexes of curvilinear multiple correlation, ρ, by the Ezekiel or short-cut methods.[18] When all relationships are additive only and not joint, these measures are satisfactory. However, when some relationships are joint, the additive measures are inadequate. The most disturbing factor in correlation analysis is the analyst's ignorance of the nature of the relationships. After having calculated a multiple correlation coefficient, the analyst still does not know whether relationships are additive or joint.

The multiple correlation coefficient for the relation of size, yields, and efficiency to income was found to be $R_{1.234} = 0.53$ (table 15, top). The partial regression coefficients in the multiple regression equation indicated the *average* rates of change in income with unit changes in each of the independent variables. There was nothing in either the correlation or regression coefficients which indicated whether the relationships were additive or joint. The ordinary procedure would be to ignore the possibility of joint relationships, unless the analyst has some logical basis to assume that they exist. In the present example, the authors knew from the tabular analysis that joint relationships were present. If they had not known this, they might have guessed it from the many studies in farm management.

---

[18] Discussed on pages 217 and 230.

TABLE 15.—RESULTS OF CORRELATION ANALYSIS OF JOINT MULTIPLE RELATIONSHIPS INVOLVING FOUR VARIABLES

RELATION OF SIZE OF FARM, X_2, CROP YIELDS, X_3, AND LABOR EFFICIENCY, X_4, TO INCOME, X_1

Additive analysis

Coefficient of multiple correlation	Coefficient of determination
$R_{1.234} = 0.53$	$R^2_{1.234} = 0.28$

Multiple regression equation

$$X_1 = +\$1.32X_2 + \$6.85X_3 + \$2.95X_4 - \$1{,}309$$

Joint analysis

Index of joint correlation	Coefficient of determination
$\rho_{1.234(\text{linear joint})} = 0.61$	$\rho^2_{1.234(\text{linear joint})} = 0.38$

Multiple regression equation

In terms of original values of X_2, X_3, and X_4

$$X_1 = -\$5.85X_2 + \$7.10X_3 + \$12.08X_4 + \$0.0515X_2X_3 - \$0.0031X_2X_4 - \$0.0987X_3X_4 + \$0.000079X_2X_3X_4 - \$951$$

In terms of deviations of X_2, X_3, and X_4 from their averages

$$X_1 = +\$0.06x_2 + \$8.05x_3 + \$3.97x_4 + \$0.0677x_2x_3 + \$0.0045x_2x_4 - \$0.0739x_3x_4 + \$0.000079x_2x_3x_4 + \$279$$

When there are three or more independent variables, correlation methods of analyzing joint relationships are very laborious and almost always impractical. Probably the simplest measure of joint correlation involving three independent variables is the mathematical index of linear joint correlation.[19] This index is calculated as follows:

1. From the three independent variables, four new independent variables were calculated to represent all the joint effects. These were

$$\begin{aligned} X_5 &= X_2X_3 &&= \text{Joint effect of } X_2 \text{ and } X_3 \\ X_6 &= X_2X_4 &&= \text{Joint effect of } X_2 \text{ and } X_4 \\ X_7 &= X_3X_4 &&= \text{Joint effect of } X_3 \text{ and } X_4 \\ X_8 &= X_2X_3X_4 &&= \text{Joint effect of } X_2,\ X_3, \text{ and } X_4 \end{aligned}$$

---

[19] Described on pages 246 to 250. The work required to calculate even this simplest measure is enormous.

2. Using all eight variables, the usual least-squares methods[20] were employed to calculate the multiple correlation coefficient $R_{1.2345678}$. This coefficient is really the index of linear joint correlation, $\rho_{1.234(\text{linear joint})}$.

The index of linear joint correlation for the relationship of size, yields, and efficiency to income was $\rho_{1.234(\text{linear joint})} = 0.61$ (table 15, bottom). The coefficient of determination, $\rho^2_{1.234} = 0.38$, indicates that joint relationships explained part of the squared variability in income not explained by the additive relationships, $R^2_{1.234} = 0.28$ (table 15, top).

Two multiple joint regression equations were calculated (table 15, bottom). One was in terms of the original values of X_2, X_3, and X_4. This was useful only in estimating the income for a given set of conditions. The other equation, which was in terms of deviations from averages, was useful chiefly for studying the effect of each independent variable with the others held constant at their average. The equation indicates that the rates of change for additive and joint effects were as follows:

	ADDITIVE	TWO-WAY JOINT	THREE-WAY JOINT
X_2	\$+0.06	$X_2X_3 =$ \$+0.0677	$X_2X_3X_4 =$ \$+0.000079
X_3	+8.05	$X_2X_4 = +0.0045$	
X_4	+3.97	$X_3X_4 = -0.0739$	

Numerous partial correlation or beta coefficients could be calculated to show the relative importance of the various joint and additive effects of the three independent variables.

COMPARISON

Both tabulation and correlation methods will show:

1. Whether joint relationships are present. However, tabular analysis shows this almost at a glance of the usual averages obtained, whereas correlation indicates joint relationships only after a very laborious process.

---

[20] The method for calculating multiple correlation coefficients is given on pages 168 to 176. There is one important pitfall away from which the student should be steered. The product moments and squared standard deviations involving X_5, X_6, X_7, and X_8 are not calculated in exactly the same manner as if those variables were not derived from X_2, X_3, and X_4. For example, the squared standard deviation of X_5 is not $\Sigma x_5^2/N$, but rather $\Sigma x_2^2 x_3^2/N$. The two quantities are not the same because AX_5, about which the deviation, x_5, is taken, is not the same as the product of AX_2 and AX_3. In terms of original values of X_2, X_3, and X_5, the squared standard deviation of X_5 would not be

$$\sigma_5^2 = AX_5^2 - (AX_5)^2$$

but

$$\sigma_5^2 = AX_5^2 - 2AX_3AX_2X_5 - 2AX_2AX_3X_5 + 4AX_2AX_3AX_5 + (AX_3)^2AX_2^2 + (AX_2)^2AX_3^2 - 3(AX_2)^2(AX_3)^2$$

2. The rate of change in income with unit changes in the independent variables.

3. The relative importance of each additive and joint effect in determining income.

Tabulation immediately shows the following fact not readily indicated by correlation: incomes for various combinations of the independent variables.

Correlation shows the following fact not indicated by tabulation: the degree of relationship between the independent variables and the dependent variable.

The only comparable measures from tabulation and correlation analysis were the rates of change. These were not in complete agreement. The average rates of change were as follows:

Additive Effects			Joint Effects		
Variables	*Tabulation*	*Correlation*	*Variables*	*Tabulation*	*Correlation*
X_2	$ −0.51	$ +0.06	X_2X_3	$ +0.0816	$ +0.0677
X_3	+6.86	+8.05	X_2X_4	+0.0109	+0.0045
X_4	+5.62	+3.97	X_3X_4	−0.0920	−0.0739
			$X_2X_3X_4$	+0.000262	+0.000079

The discrepancies were due to differences between the averaging methods of tabulation and correlation analysis.

In the equations of relationship by the two methods, many of the above discrepancies seem to be compensating. Based on these equations, estimates of income for values of the independent variables well within the range of most of the actual data were as follows:

Combinations			Income Estimated from Equations Determined by	
Size, X_2	*Yields*, X_3	*Efficiency*, X_4	*Tabulation*[21] *analysis*	*Correlation*[22] *analysis*
Small, 200	poor, 80	low, 140	$ − 96	$ − 51
Small, 200	good, 120	low, 140	+130	+180
Small, 200	average, 100	average, 200	+264	+256
Large, 400	poor, 80	high, 260	+381	+348
Large, 400	good, 120	high, 260	+805	+758

The differences in the incomes estimated by the two methods for the five combinations of independent variables were small, ranging from $8 to $50. For combinations of the independent variables not commonly existing in this community, the two equations give widely different estimates of income, as follows:

---

[21] Page 289. [22] Page 292.

Combinations			Income Estimated from Equations Determined by	
Size, X_2	*Yields, X_3*	*Efficiency, X_4*	*Tabulation analysis*	*Correlation analysis*
Very large, 1,000	very poor, 60	very low, 100	\$ − 3,477	\$ −2,500
Very large, 1,000	very good, 200	very high, 400	+10,855	+6,922

A combination of very large farms with very poor yields and very low efficiency is a very uncommon occurrence. Both methods gave the expected negative estimated incomes, but the difference between the two estimates was \$977. A combination of very large farms, very good crop yields, and very high efficiency is also an uncommon occurrence. Both methods gave the expected high positive incomes. However, the difference, \$3,933, was rather large. Neither equation of relationship is accurate in estimating incomes on farms with characteristics widely different from those of the main body of the farms studied.

When relationships are joint, correlation analysis is much more laborious than tabular analysis. This is true regardless of whether tabulating and calculating machines are available. This fact can be appreciated only by those who have applied both methods. In this chapter, much more space has been devoted to tabulation than to correlation analysis of joint relationships. This is not indicative of the relative work involved in the two methods. Two things must be kept in mind:

1. Practically all the work of tabular analysis was shown in detail, while practically none of the extensive and intricate calculations of correlation analysis were shown.

2. The tabular analysis was carried farther than necessary, merely to show that results could be obtained which were comparable to the results of correlation. For practical purposes, the simpler steps of tabular analysis are sufficient.

The work involved in correlation analysis of joint relationships is so laborious that this type of approach is usually ruled out as impracticable. For the analyst, the choice usually lies between additive correlation analysis, which does not take into account joint relationship, and tabular analysis, which does.

SIMPLICITY OF METHODS

From the standpoint of simplicity of the method of analysis, tabulation has the overwhelming advantage over correlation. In tabulation, all the calculations are simple, consisting of only a few additions, divisions, and subtractions of relatively simple numbers. In linear correlation, the calculations are long and involved, dealing with numerous products, squares, simultaneous equations, square roots, and the like. In curvilinear correlation, the amount of work is still greater. In analyzing joint relationships, the work involved in correlation is tremendous, and the advantage of tabular analysis is overwhelming.

Simplicity of method is extremely important for at least two reasons:

1. It means a great saving of time.
2. It makes possible the analysis of relationships by the great mass of research workers. The analyst can visualize a simple process, and, because he understands it, he has confidence in it and therefore uses it.

FLEXIBILITY OF METHODS

Tabulation is the more flexible type of analysis. The nature of the relationship is not assumed at the outset. For this reason, the nature of the relationship need not be known prior to the analysis. With tabular analysis, the technique is the same whether the relationships are linear or curvilinear, additive or joint. The discovery of these characteristics comes in the interpretation after the tabulations have been made. With correlation analysis, the techniques depend directly on whether the relationships are linear or curvilinear and whether additive or joint. A different method is used for each different type of relationship. However, the real difficulty lies in the worker's ignorance of the relationship prior to making the analysis. For this reason, the worker who uses correlation may make many false starts before finding the correct method.

NUMBERS OF OBSERVATIONS

In dealing with a large number of observations, tabulation is usually preferable to correlation because of its simplicity and flexibility. For small numbers of observations, the usefulness of both methods is somewhat limited, but tabular analysis becomes less useful than correlation.

Tabular analysis is "wasteful" of data. In comparing the group averages of a table, the reliability of the comparison is limited by the smallest group. Since the subgroups in a table are rarely equal, some information is almost always wasted. On the other hand, correlation analysis does not "waste" any data. In averaging a relationship, it makes more efficient use of all the items. In correlation analysis, nothing is wasted by comparing unequal groups. Average relationships are approximated by comparing each item with each other item.

When the number of observations is large, the greatest advantages lie with the tabular method. When the number of items is small, this disadvantage of tabulation outweighs all its advantages.

NON-NUMERICAL VARIABLES

When one or more independent variables are expressed in non-numerical terms, relationships cannot be analyzed with the usual methods of correlation. On the other hand, tabular methods are just as simple and effective for non-numerical as for numerical independent variables.

SIMPLICITY OF RESULTS

From the standpoint of simplicity of presentation and interpretation of the results by the layman, tabulation again has the overwhelming advantage. The results of tabulation are usually simple averages widely understood. The results of correlation are abstract correlation coefficients and the equally perplexing concrete rates of change. It is true that a regression equation may be a concise description of the nature of the relationship and that a correlation coefficient is a concise measure of the degree of relationship. However, such concise descriptions are useless to those who do not understand their meaning. The rank and file of research workers and the laymen understand a table but are confused by coefficients.

Since relationships are often very complex, a large amount of judgment and common sense is required in their analysis and interpretation. This applies equally well whether tabulation or correlation methods are used.

SUMMARY

The relative usefulness of tabulation and correlation analysis lies in the adequacy of results and the ease with which those results may be obtained, interpreted, and presented.

On the basis of adequacy of results, each method has its peculiar advantages and disadvantages, as follows:

1. Since tabular analysis makes no assumptions as to the nature of relationships, it usually shows relationships as they exist. On the other hand, linear correlation methods show curvilinear correlation incorrectly, and additive correlation methods show joint correlation incorrectly.

2. When there are marked interrelationships among independent variables, tabular analysis may lead to somewhat erroneous conclusions concerning the relative effects of the interrelated variables.

3. Correlation methods show the degree of association between the independent and dependent variables, whereas tabular analysis gives no answer to this question.

Other types of results, such as direction of relationship and rates of change are obtained with about equal accuracy by either method, provided that the correct procedure has been chosen.

There are greater differences between tabulation and correlation from the standpoint of the ease of obtaining, interpreting, and presenting the results than from the standpoint of the results themselves. Tabular analysis requires much less time than correlation. The results are in more simple terms and are more easily understood than the results of correlation. The results of tabulation are usually in a form in which the layman can understand them. This cannot be said of the results of correlation analysis.

Tabular analysis has such important advantages over correlation that it is by far the more desirable method when the number of observations is large. Simplicity of technique and simplicity of interpretation are of greatest practical importance and explain the almost universal acceptance of tabular analysis as a tool for scientific research.

When the number of observations is small, tabular analysis has an important defect. The amount of data represented by group averages is too small to give reliable results.

Correlation is merely a substitute for tabulation when the number of observations is too small to give reliable group averages. Correlation, like tabulation, is an averaging process. Correlation logically applies to relationships which cannot be studied by tabulation because of insufficient observations. Its real service to the research worker is in using all the items to reveal relationships which otherwise could not be observed.

The choice of tabular or correlation analysis is usually determined by the quantity of data available. Ordinarily, the two methods do not compete. At the dividing line, where the data are neither scanty nor numerous, the student must make the choice. The dividing line is not distinct. It depends on the number of groups in the tabulation, the distribution of the observations, the degree of relationship, the amount of variability in the data, interrelationships among independent variables, whether relations are joint or additive, and the like. When in doubt, the student might employ both methods. Experience is the best teacher in selecting the more suitable method for a problem to which neither method is unquestionably adapted.

The relative importance of tabulation and correlation methods is highly distorted in statistical courses and textbooks. Much is said about correlation, but tabulation seldom receives even passing mention. One measure of relative importance is the number of studies in which the two methods were used. Research literature overwhelmingly emphasizes the tabulation approach. Textbooks, however, emphasize correlation, more because of its difficulty and its mathematical precision than because of its importance. Tabulation, as a method of analyzing relationships, is omitted from textbooks because it is so simple, although its simplicity is the chief reason for its importance.

CHAPTER 16

MEASURES OF RELIABILITY

Many of the doctrines in this world are not true. Whether a belief is true or untrue depends partly on the method by which that belief was developed. Most persons arrive at a given conviction by theory, observation, or measurement. Of all the doctrines developed by theory, only a small percentage proves to be true. Of all the things observed, a high percentage proves to be fact. An even higher proportion of doctrines based on measurement proves to be true.

Since time and expense restrict the application of measurement to an insignificant percentage of the world's problems, observation is man's primary method of determining truths.

Most accepted beliefs are products of coincidences of events. The establishment of truth by coincidence is complicated by man's ability to distinguish between usual and unusual, that is, recurring or nonrecurring coincidences. The observation of usual, recurring coincidences ordinarily results in convictions of which a high percentage is true. Superstition develops from observation and generalization from unusual and rarely recurring coincidences. "Coincidences, in general, are great stumbling-blocks in the way of that class of thinkers who have been educated to know nothing of the theory of probabilities."[1]

The subject of statistics concerns *measurement*—the measurement of various characteristics of phenomena, such as central tendency, dispersion, and relationships. The principles developed by measurement are more often correct than those synthesized by other methods. However, in the search for truth, perfection is not attained even in measurement. Coincidence is the curse of measurement as well as of observation. The difference is only in degree.

A fact is what it is for several different reasons. The force or condition being studied may or may not be one of those reasons. There may be a coincidence due to chance alone. "Chance" includes the effects of all the factors not considered. Of the many factors not considered, some may operate in such a way that the fact appears to be the result of the force being studied. This is what happens when a coincidence occurs. One of the important problems of measurement is to determine how much of the measure is due to the factors under consideration and how much is due to chance or the factors not considered.

---

[1] The Murders in the Rue Morgue.

The need for measures of reliability arises from lack of complete information. One coincidence is almost meaningless. After a few repetitions, convictions are formed. If the coincidences become numerous, the convictions become accepted facts. Part way between the single coincidence and the established fact lies a field of doubt with varying degrees of uncertainty. Determining the point at which an "apparent" fact becomes a "true" fact is the statistical problem of measuring reliability.

In the measurement of phenomena, information is almost never complete. The statistician is forced to work with samples. The collection and analysis of complete data are almost always physically impossible. The statistician must use the sample, a small number of observations, with which to generalize about a "universe," all the possible observations.

This process of estimating the characteristics of the universe from those of a sample is called statistical induction or inference. It is one of the most important but most dangerous steps in statistical analysis. Mills likens this step to a leap in the dark.

Because data vary, it is known in advance that the apparent facts about the universe shown by the sample are probably not absolutely true. Measures of reliability indicate the range within which the observed and true facts have a given probability of agreement. Conversely, they also indicate the probability that the observed and true facts agree within a certain range.

A sample should be representative of the universe which it describes. The process of statistical inference is, of course, based on this assumption. Obtaining a representative sample almost always meets with practical difficulties in the form of biases. A bias is some kind of prejudice operating in such a way that the sample does not show the same characteristics as the universe. When a sample is biased, the differences between sample and universe are not all due to chance fluctuations. Bias may be of many types. Examples of biases are tendencies to overstate yields or understate assessments; to interview the more desirable families, farms, or stores. The degree of bias depends on the degree to which the sample is a random selection. The problem of obtaining a representative sample simmers down to a sufficiently random method of selection.

Approximately random samples, let alone purely random samples, are rarely, if ever, obtained in research work. "Random sample" is a fascinating expression with which statistical theorists love to play. It has relatively little practical significance.

If a sample is known to be biased, correction should be made for the bias. If it is not known whether bias is present, that bias itself becomes a chance fluctuation, and there is nothing to do about it except to conclude that the sample is representative.

Some research workers do not use measures of reliability, contending that they are not applicable to their problems. These measures are not necessary when no generalizations are to be made or when there is an infinite number of observations. The latter condition never exists; and the former, rarely. Measurement is worthless unless generalizations are made from the results.

Some workers ignore measures of reliability because they assume that their data include the whole population or universe. Measurement of the whole possible population is almost never accomplished. Data which seem to include a universe actually include only a part of a larger universe. All the farms in Webster County, Iowa, are "all the farms," "the universe," "the total population." They are only a part, however, of a larger population, Iowa, which, in turn, is only a part of a still larger universe, the Corn Belt. Since Webster County is a population in itself, no generalization concerning Webster County would be needed. The facts about Webster County would be at hand. However, the one universe, Webster County, would probably be used to generalize concerning other parts of Iowa or the Corn Belt. In that case, Webster County is a sample from a universe or population.

A universe or population is all the things within any prescribed limits. Nearly every group of data with which the student works is usually thought of as a sample. In another sense, it is a universe.

Some workers do not use measures of reliability for time series. They claim that the Chicago price of No. 3 oats from 1920 to 1940, for example, is a total population and that the measures of reliability do not apply. In a sense, this period of time is a small sample of a universe, "eternity." A population whose individuals are different units of time is subject to elements of change that are not present among data all relating to the same period of time. These disturbing elements probably increase the unreliability of time series. The usual measures of reliability often overstate the significance of the conclusions drawn from time series. Another difference in time series is that the observations usually follow in definite sequence and are sometimes dependent on those preceding. Whatever the differences between time series and other samples, the measures of reliability for the latter should also be used for the former in the absence of more suitable measures.

In the scheme of determining what are facts, measures of reliability supplement the measures of description. Measures of reliability indicate the degree of certainty to be attached to the description.

CHAPTER 17

STANDARD ERRORS

All things fluctuate from a multiplicity of causes. Some of these fluctuations are so small that they are not visible to the naked eye; for example, the length of a ruler, which is assumed to be fixed, in reality varies from time to time. Other fluctuations are very violent, for example, size of profits and numbers of bacteria.

The degree of variability among observations has been universally measured by the standard deviation.[1]

NORMAL FREQUENCY CURVES

The nature of variability is indicated by frequency curves. Frequency curves tend to bell shapes, indicating a concentration of observations at a central point, with diminishing numbers on either side. When the distribution is "normal," there are definite relationships between the standard deviation and the frequency curve. A range of one standard deviation either side of the mean, the high point on the curve, includes 68.27 per cent of the observations. Ranges of two and three standard deviations include 95.45 and 99.73 per cent of the observations, respectively. However, most distributions of observations are not normal, and these percentages do not apply. The difference between these percentages and those that do apply depends on the degrees of skewness and kurtosis.

VARIABILITY IN AVERAGES

Up to this point, only the variability in individual observations has been studied.[2] A statistical measure such as the arithmetic mean also fluctuates. A large body of data, such as a "population," has only one arithmetic mean. When it is assumed that 1,000 cases of eggs in New York City are a universe or population, the arithmetic mean of the weights for that population is 40.31 pounds (table 1). However, if the population is divided into a number of samples, the arithmetic means of those samples will not be the same as the mean of the population, or the same as one another. The 1,000 cases of eggs were divided into 20 lots or samples of 50 cases each. The average weights of the 20 lots ranged from 38.80 to 41.94 pounds (table 1). Only 2 lots (5 and 15) averaged the same, and no one lot averaged exactly the same as the whole population of 1,000 cases.

---

[1] The standard deviation is sometimes called the standard error of an observation.

[2] The standard deviation, $1,245, measured the variability in incomes on *individual* farms about the arithmetic mean of incomes, $1,212 (table 6, page 46).

TABLE 1.—THE STANDARD DEVIATION IN A "POPULATION" AND IN ITS "SAMPLE" MEANS

WEIGHTS OF EGGS PER CASE FOR 20 LOTS OF 50 CASES EACH
NEW YORK CITY, JANUARY 1931

Sample number	Number of cases	Average weight, pounds	Population of individual cases
1	50	41.14	$Ma = 40.31$ $\sigma = \sqrt{\frac{27{,}457}{1{,}000}}$
2	50	41.50	
3	50	39.32	$= 5.24$
4	50	39.88	
5	50	39.80	Samples of 50 cases each
6	50	40.58	
7	50	40.34	$Ma = 40.31$
8	50	41.94	$\sigma = \sqrt{\frac{11.2788}{20}}$
9	50	38.80	
10	50	40.26	$= 0.75$
11	50	39.62	
12	50	39.68	
13	50	40.32	Sample 1 : 50 individual cases
14	50	40.44	$Ma = 41.14$
15	50	39.80	
16	50	40.42	$\sigma = \sqrt{\frac{1{,}171}{50}}$
17	50	39.72	$= 4.84$
18	50	40.64	
19	50	41.32	
20	50	40.62	$s = \sqrt{\frac{1{,}171}{50 - 1}}$
Total or average	1,000	40.31	$= 4.89$

The variability in the sample means can be measured by the standard deviation in those means. This standard deviation is calculated in the same general manner as the standard deviation in individual observations.

For individual observations,

$$\sigma = \sqrt{\frac{\Sigma(X - AX)^2}{N}}$$

For arithmetic means,

$$\sigma_{Ma} = \sqrt{\frac{\Sigma(X - AX)^2}{N}}$$

where σ = standard deviation in individual observations.
X = individual observations.
AX = average of individual observations.
N = number of observations.

where σ_{Ma} = standard deviation in arithmetic means of samples.
X = arithmetic mean of a sample.
AX = arithmetic mean of all samples.
N = number of samples.

To find the standard deviation in the individual observations, the deviation of the weight of each of the 1,000 cases from their mean, 40.31 pounds, was calculated and squared. The sum of the 1,000 squared deviations, 27,457, was divided by the number of individual observations, 1,000. The quotient, 27.457, is the squared standard deviation, and its square root, 5.24, is the standard deviation in the population of 1,000 individual cases of eggs.

To find the standard deviation in the means of the 20 samples, the deviation of each sample mean from the mean of the 20 sample means, 40.31 pounds, was calculated and squared. The sum of the 20 squared deviations, 11.2788, was divided by 20. The quotient, 0.564, is the squared standard deviation, and its square root, 0.75, is the standard deviation in the 20 sample means.

To distinguish between the standard deviation in observations, $\sigma = 5.24$, and the standard deviation in the means, $\sigma_{Ma} = 0.75$, the former is commonly called plain standard deviation; and the latter, the *standard error of the mean.*

STANDARD ERROR OF THE MEAN

The standard error of the arithmetic mean has a definite relationship to the standard deviation. Because of this relationship, the variability in means can be estimated from the variability in individual observations.

$$\sigma_{Ma} = \frac{\sigma}{\sqrt{N}}$$

where σ_{Ma} is the standard error of the mean, σ is the standard deviation of the population, and N is the number of observations in the samples. The standard deviation in the sample means was 0.75 by calculation; and by estimation,

$$\sigma_{Ma} = \frac{5.24}{\sqrt{50}} = \frac{5.24}{7.07} = 0.74$$

The standard error of the mean, $\sigma_{Ma} = 0.75$ or 0.74, has the same relationship to a normal distribution of means that the standard deviation, $\sigma = 5.24$, has to a normal distribution of individual observations.

A range of 1 standard error either side of the mean of sample means includes 68.27 per cent of the sample means. Ranges of 2 and 3 standard errors include 95.45 and 99.73 per cent, respectively.

There is an important difference between distributions of series of individual observations and distributions of sample means. Distributions of individual observations are rarely near normal. They are usually either badly skewed or too peaked or flat-topped. They may even be U-shaped or J-shaped. On the other hand, distributions of means tend to normality quite closely. Even when the distribution of individual items is U-shaped, the distribution of the means of samples from such a series will tend to be normal.

The use of the standard deviation in describing distributions of individual items is limited by the failure of such distributions to approximate normal curves. Standard errors of the means, σ_{Ma}, can be used to interpret a distribution of means because these distributions are almost always approximately normal.

If one knew the average weight of the 1,000 cases of eggs, 40.31 pounds, and the standard error of the means of samples with 50 cases each, $\sigma_{Ma} = 0.75$, one could predict, within certain limits, the weight of a sample of 50 cases. About two-thirds, 68.27 per cent, of such samples would weigh within 0.75 pound of the population average weight, 40.31 pounds. This range would be from 39.56 to 41.06 pounds (40.31 ± 0.75). The probability that any one sample would weigh within this range would be 0.6827, or 2 chances out of 3. Fourteen of the 20 sample means of egg weights, 70 per cent, actually were within this range (table 1). Likewise, the probability that the sample would weigh within 2 standard errors from the mean, $Ma \pm 2\sigma_{Ma}$, would be 0.9545; and within $3\sigma_{Ma}$, 0.9973. However, this use of the standard error is of little value in practical problems because:

(*a*) The average weight for the population is rarely known.

(*b*) If the average weight for the population were known, there would be no object in examining samples from that population.

(*c*) If a sample *were* examined, its average weight could be determined more accurately and easily by direct calculation from the sample than by estimation from the population.

GENERALIZING FROM A SAMPLE

The problem of statistical inference is generalizing *from* the sample *to* the population. The standard error of the mean is not valuable for estimating the mean of the sample from the population, but *is* valuable in estimating the mean of the population from the sample. The probability was 0.68 that the mean of a sample of 50 cases of eggs would fall within a range of 1 standard error either side of the population mean. It follows that the probability is 0.68 that the mean of the population would fall within a range of 1 standard error either side of the mean of that sample. For egg weights, 14, or 70 per cent, of the sample means were within 1 standard error of the population mean. It follows that the population mean was within 1 standard error of the means of 14 samples.

If a person wished to examine the weights of eggs in the population of 1,000 cases, he might weigh only the first 50 cases. The average weight, 41.14 pounds per case, would be the best estimate of the average weight of the whole population. However, it would not be an exactly correct estimate. The differences which might occur between the sample and population means and the chances of the differences occurring would be indicated by the standard error of the mean. The stumbling-block here is that one would not know the standard error of the mean of a sample without first having obtained the standard deviation of the population. In the formula $\sigma_{Ma} = \sigma/\sqrt{N}$, the standard deviation, σ, is for the population rather than for the sample. The impossibility or impracticability of calculating σ from the population is circumvented by estimating it from the sample as follows:

$$\sigma_{(\text{population})} = \sqrt{\frac{N}{N-1}}\,\sigma_{(\text{sample})}$$

where σ (sample) is the standard deviation in the individual observations in the sample calculated in the usual manner according to the formula $\sigma = \sqrt{\Sigma(X - AX)^2/N}$. The standard error of the mean based on the standard deviation in the observations of a sample[3] would be $\sigma_{Ma} = \dfrac{\sigma(\text{sample})}{\sqrt{N-1}}$.

The estimate of the standard deviation of the population from the sample is usually denoted by s and calculated as follows:

$$s = \sqrt{\Sigma(X - AX)^2/(N - 1)}.$$

In terms of this population estimate, the standard error of the mean would be calculated as follows: $\sigma_{Ma} = \dfrac{s}{\sqrt{N}}$.

---

[3] Since $\sigma_{Ma} = \dfrac{\sigma_{(\text{population})}}{\sqrt{N}}$, and $\sigma_{(\text{population})} = \sqrt{\dfrac{N}{N-1}}\,\sigma_{(\text{sample})}$, then $\sigma_{Ma} = \dfrac{\sqrt{\dfrac{N}{N-1}}\,\sigma_{(\text{sample})}}{\sqrt{N}} = \dfrac{\sigma_{(\text{sample})}}{\sqrt{N-1}}$.

The standard deviation of sample 1 of the egg weights was 4.84 pounds (table 1). The standard deviation of the population estimated from sample 1 was 4.89 pounds (table 1). Based on sample 1, the standard error of the mean was

$$\sigma_{Ma} = \frac{\sigma}{\sqrt{N-1}} = \frac{4.84}{\sqrt{50-1}} = 0.69 \text{ pound}$$

or

$$\sigma_{Ma} = \frac{s}{\sqrt{N}} = \frac{4.89}{\sqrt{50}} = 0.69 \text{ pound}$$

If it is assumed that $\sigma_{Ma} = 0.75$ for the population, the estimate from sample 1, $\sigma_{Ma} = 0.69$, is fairly accurate. Of course, in this particular example, the estimate is rather useless because the standard error from the population is known.[4] However, in most problems of sampling, the characteristics of the population are not known.

Knowing only the average weight for sample 1 and the standard error of the mean estimated from sample 1, $Ma = 41.14$ pounds and $\sigma_{Ma} = 0.69$, one can guess the average for the population. The chances are 68 out of 100 that the population mean lies within a range of 1 standard error, 0.69, of the sample mean, 41.14. This range is from 40.45 to 41.83. The chances are about 95 out of 100 that the population mean lies within a range of 2 standard errors, 39.76 to 42.52. Actually, the range of 1 standard error did not include the population mean, 40.31, but a range of 2 standard errors did include it. If the student had deduced that the population mean was within 1 standard error of the sample mean, he would have been wrong, even though the odds would have been 68 to 32 in his favor.

Any estimates concerning a population that are made from a sample can be stated only in terms of probabilities. There is always some chance of being wrong. When a person states that the population mean is within 1 standard error of the sample mean, there are 32 chances out of 100 that he is wrong. As the range of the estimate increases, the chances of being wrong decrease. If the population mean is estimated to be within 3 standard errors of the sample mean, the chances of being wrong are less than 1 per cent.

The use of the standard error of the mean is illustrated by the estimation of milk production in Wisconsin. Canvassing the state to obtain the production of every herd would be impracticable. However, it would be fairly easy to visit a few herds, say 300. Assume that the 300 herds averaged 5,500 pounds of milk per cow, and the standard deviation among these herds was 1,000 pounds. The standard error of the mean would be $\frac{1,000}{\sqrt{300 - 1}} = \frac{1,000}{17.3} = 58$ pounds per cow. The range of 3 standard errors from the mean, 5,500, would be 5,326 to 5,674 pounds. The chances would be 99.7 out of 100 that production for the state of Wisconsin was within this range. Likewise, the chances would be 95.45 out of 100 that the average for Wisconsin was between 5,384 and 5,616, a range of 2 standard errors either side of the mean, 5,500.

---

[4] Strictly speaking, the population, 1,000 cases, is itself a sample of a larger universe. The standard error of the means of samples of 1,000 cases could be calculated.

These estimates with the accompanying probabilities would be valid only provided that the assumptions in sampling were fulfilled. The sample of 300 herds must be representative of Wisconsin dairy farms.

TABLE 2.—A COMPARISON OF STANDARD ERROR AND PROBABLE ERROR

Amounts of error	Probability of population mean occurring within range of the error of sample mean	Approximate chances of occurrence	Degrees of probability
Standard errors			
$\pm 0.6745\,\sigma$	0.5000	1 to 1	Equal, fifty-fifty
$\pm 1.0000\,\sigma$	0.6827	2 to 1	Favorable
$\pm 2.0000\,\sigma$	0.9545	21 to 1	High
$\pm 3.0000\,\sigma$	0.9973	369 to 1	Practical certainty
Probable errors			
±1 P.E.	0.5000	1 to 1	Equal, fifty-fifty
±2 P.E.	0.8227	5 to 1	Favorable
±3 P.E.	0.9570	22 to 1	High
±4 P.E.	0.9930	142 to 1	Practical certainty
±5 P.E.	0.9993	1,340 to 1	Practical certainty

STANDARD ERROR AND PROBABLE ERROR

Some statisticians like to think in terms of probable errors rather than standard errors. A probable error is 0.6745 times the standard error. One probable error either side of the arithmetic mean of the population includes one-half the means of the samples. There are 50 chances out of 100 that the mean of the population will be within 1 probable error of the sample mean. Likewise, there are 50 chances that it will fall outside this range. Roughly, 3 probable errors equal 2 standard errors (table 2).

Since probable and standard errors have a constant relationship, their interpretation is essentially the same. The student may use whichever measure appeals to him. However, the probable error has been passing out of general use.

STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO MEANS

The standard error of the difference between two means is very important. Means, or simple averages, are the most widely used statistical tools. Comparing one average with others is the most common method of using such averages. The comparison of averages is the basis of the tabular method of analyzing relationships.

Comparison of averages is equivalent to observing differences between averages. One of the most bothersome problems of comparison is whether differences are large enough to be significant, that is, large enough to be considered not due to chance. Most persons guess whether differences are significant. Statistical technique has been developed to test accurately the significance of differences. The standard error of the difference between two means is given by

$$\sigma_{D_{Ma}} = \sqrt{\sigma^2_{Ma_1} + \sigma^2_{Ma_2}}$$

where D_{Ma} = difference between Ma_1 and Ma_2, σ_{Ma_1} = standard error of the mean of one series, and σ_{Ma_2} = standard error of the mean of a second series.

The average yields of two varieties of corn on 10 and on 20 farms in the same locality were 44.4 and 54.7 bushels per acre, respectively. The difference in the yields was 10.3 bushels. The problem is to test the likelihood and amount of the difference between these two varieties for the whole locality. For this purpose, the standard error of the difference is obtained. Where $\sigma^2_{Ma_1} = 2.69$ and $\sigma^2_{Ma_2} = 1.91$,

$$\sigma_{D_{Ma}} = \sqrt{2.69 + 1.91} = \sqrt{4.60} = 2.1 \text{ bushels of corn}$$

The interpretation of the standard error of the difference between the two means is the same as for the standard error of the mean. A range of 1 standard error, 2.1, on either side of the actual difference, 10.3 bushels, would be from 8.2 to 12.4. The probability is 0.68 that the difference between the yields of the two varieties would be between 8.2 and 12.4 on all farms in this locality with the same conditions. Likewise, the probability is 0.997 that the difference lies between 4.0 and 16.6, a range of 3 standard errors on either side of the observed difference, 10.3.

Another method of interpreting the standard error of the difference between two means is:

1. Assume that no difference existed between the two varieties on all farms in the locality.

2. Calculate the deviation of the observed difference from zero in terms of standard errors.

3. The closer the observed difference to zero, that is, the smaller the deviation in terms of standard errors, the less likelihood there is that a difference did exist in the two varieties for the whole locality.

The difference, 10.3, for the 30 farms was a deviation of 4.9 standard errors from zero [$(10.3 - 0) \div 2.1 = 4.9$]. The probability would be very high that the difference between the two varieties in the whole locality would be between 0 and 20.6, a range of 4.9 standard errors either side of 10.3 bushels. Stated another way, it would be practically certain that some difference did exist between the two varieties throughout the whole locality.

The standard error of the difference may also be calculated from the standard deviations of the two series.

$$\sigma_{D_{Ma}} = \sqrt{\frac{\sigma_1^2}{N_1 - 1} + \frac{\sigma_2^2}{N_2 - 1}}$$

where σ_1 and σ_2 are the standard deviations of the individual observations in the first and second series, respectively, and N_1 and N_2 are the numbers of observations in those two series. The standard error of the difference by this method would be identical to that determined above, 2.1.

In the above formula, the standard deviations of yields of both varieties are used. Often, the two standard deviations are *pooled* into one estimate of the population standard deviation according to the following formulas:

A

$$s_p = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2}}$$

B

$$s_p = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}$$

For the corn-yield problem, the following facts are known:

$Ma_1 = 44.4$ $\quad Ma_2 = 54.7$ $\quad D_{Ma} = 10.3$

$N_1 = 10$ $\quad N_2 = 20$ $\quad \sigma_1^2 = 24.2$ $\quad \sigma_2^2 = 36.3$

$\Sigma x_1^2 = 242$ $\quad \Sigma x_2^2 = 726$ $\quad \sigma_{Ma_1}^2 = 2.69$ $\quad \sigma_{Ma_2}^2 = 1.91$

The pooled estimate of the standard deviation may be obtained as follows:

A

$$s_p = \sqrt{\frac{(10 \times 24.2) + (20 \times 36.3)}{10 + 20 - 2}} = \sqrt{\frac{968}{28}} = 5.88$$

B

$$s_p = \sqrt{\frac{242 + 726}{10 + 20 - 2}} = \sqrt{\frac{968}{28}} = 5.88$$

When the *pooled* standard deviation is used, the standard error of the difference is

$$\sigma_{D_{Ma}} = s_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$\sigma_{D_{Ma}} = 5.88\sqrt{\frac{1}{10} + \frac{1}{20}} = 5.88\sqrt{0.1 + 0.05} = 5.88(0.387)$$

$$= 2.3$$

The standard error of the difference using the pooled standard deviation, 2.3, was slightly larger than that using the standard errors of the two means, 2.1. The advantage of using the pooled standard deviation is that it weights the two series according to the number of observations. There were more farms with variety two than with variety one. Furthermore, yields for variety two were more variable than for variety one, $\sigma_2^2 = 36.3$ and $\sigma_1^2 = 24.2$. Consequently, the standard error of the difference was greater when the variability was weighted than when unweighted.

Standard errors can be applied to many other statistical measures. However, because these measures are of lesser importance, the formulas for their standard errors are set forth with little explanation.

STANDARD ERROR OF SECOND DIFFERENCES

The standard error of the difference between differences between means can be calculated. For instance, if four means are 1, 3, 6, and 11, two differences, 2 and 5, may be compared (3 − 1 and 11 − 6). The standard error of the second difference, 3, may be calculated. The formula is

$$\sigma_{D_1-D_2} = \sqrt{\sigma_{D_1}^2 + \sigma_{D_2}^2}$$

where σ_{D_1} is the standard error of one difference between means; and σ_{D_2}, the standard error of the other difference between means; or

$$\sigma_{D_1-D_2} = \sqrt{\sigma_{Ma_1}^2 + \sigma_{Ma_2}^2 + \sigma_{Ma_3}^2 + \sigma_{Ma_4}^2}$$

or

$$\sigma_{D_1-D_2} = s_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + \frac{1}{N_4}}$$

where s_p is the pooled estimate of the standard deviation of the population.

An application of standard errors of second differences is given on pages 332 and 335.

STANDARD ERRORS OF FREQUENCIES AND PROPORTIONS

The standard error of a class frequency is given by

$$\sigma_f = \sqrt{\frac{Nf - f^2}{N - 1}} \quad \text{or} \quad \sqrt{\frac{f(N - f)}{N - 1}}$$

where f = the frequency, and N = the number of observations in the total distribution. The standard error of the difference between two class frequencies[5] is

$$\sigma_{D_f} = \sqrt{\sigma_{f_1}^2 + \sigma_{f_2}^2}$$

or more accurately from a sample,

$$\sigma_{D_f} = \sqrt{\frac{2(Nf_o - f_o^2)}{N - 1}}$$

where $f_o = \frac{f_1 + f_2}{2}$, and N = the observations in the total distribution. When there are only two frequencies in the total distribution, the standard error of their difference is

$$\sigma_{D_f} = \sqrt{\frac{N^2}{N - 1}}$$

where N = the observations in the total distribution.

An application of standard errors of frequencies and differences between frequencies is given on pages 339 to 341.

The standard error of a proportion is given by

$$\sigma_p = \sqrt{\frac{f(N - f)}{N^2(N - 1)}} \quad \text{or} \quad \sqrt{\frac{pq}{N - 1}}$$

where f = the frequency from which the proportion is calculated, N = the observations in the total distribution, p = the proportion, and $q = 1 - p$.

The testing of differences between proportions is really two problems. There are differences: (*a*) between two proportions in the same distribution, and (*b*) between two proportions in two different distributions. The student must distinguish between the two types of differences because their standard errors are not the same.

(*a*) The standard error of the difference between two proportions in the same frequency distribution is given by

$$\sigma_{D_p} = \sqrt{\frac{2p_oq_o}{N - 1}}$$

---

[5] Assuming that the two frequencies are in the same total distribution containing three or more frequencies.

where p_o is the average of the two proportions, q_o is $1 - p_o$, and N is the number of observations in the total distribution. When there are only two proportions in the whole distribution, the standard error of their difference is simply

$$\sigma_{D_p} = \sqrt{\frac{1}{N-1}}$$

(*b*) The standard error of the difference between two proportions in two different frequency distributions is given by

$$\sigma_{D_p} = \sqrt{p_o q_o\left(\frac{1}{N_1 - 1} + \frac{1}{N_2 - 1}\right)}$$

where p_o = the weighted average of the two proportions, $q_o = 1 - p_o$, N_1 = the observations in one distribution, and N_2 = the observations in the other distribution.

Standard errors of proportions and their differences are not reliable when the proportions are very small, near 0, or very large, near 1.0. The interpretation of these standard errors for less extreme proportions is the same as for averages.

Applications of the standard error to (*a*) differences between proportions in the same distribution and (*b*) differences between proportions in different distributions are given on pages 340 and 341.

STANDARD ERRORS OF OTHER STATISTICAL MEASURES

The standard error of the median is greater than that for the mean,

$$\sigma_{M_e} = 1.2533\frac{\sigma}{\sqrt{N-1}} \quad \text{or} \quad 1.2533\frac{s}{\sqrt{N}}$$

Apparently, the chances of estimating the median within a certain range are less than for the arithmetic average. The standard error of the difference between two medians may be obtained by two methods:

$$\sigma_{D_{M_e}} = \sqrt{\sigma^2_{M_{e1}} + \sigma^2_{M_{e2}}}$$

or

$$\sigma_{D_{M_e}} = 1.2533 s_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

where s_p is the pooled estimate of the standard deviation for the population calculated according to the formula, $s_p = \sqrt{(\Sigma x_1^2 + \Sigma x_2^2)/(N_1 + N_2 - 2)}$. The interpretation of standard errors of medians and differences between medians is the same as for arithmetic means.

The standard errors of the first and third quartiles are

$$\sigma_{Q_1} = \sigma_{Q_3} = 1.3626\frac{\sigma}{\sqrt{N-1}} \quad \text{or} \quad 1.3626\frac{s}{\sqrt{N}}$$

The standard error of the difference between two first quartiles[6] (or two third quartiles) is

$$\sigma_{DQ} = \sqrt{\sigma_{Q_1}^2 + \sigma_{Q_1'}^2}$$
$$= 1.3626 s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

In general, the standard errors of measures of dispersion are less accurate than those of central tendency. When the parent population is normally distributed, the standard error of the standard deviation is given by

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2(N-1)}} \quad \text{or} \quad \frac{0.7071\sigma}{\sqrt{N-1}} \quad \text{or} \quad \frac{0.7071s}{\sqrt{N}}$$

A more accurate formula for all types of distributions is given by $\sigma_\sigma = \sqrt{\frac{\mu_4 - \sigma^4}{4N\sigma^2}}$, where $\mu_4 = \frac{\Sigma x^4}{N}$. The standard error of the difference between two standard deviations may be obtained by two methods:

$$\sigma_{D\sigma} = \sqrt{\sigma_{\sigma_1}^2 + \sigma_{\sigma_2}^2}$$

or, using pooled standard deviation,

$$\sigma_{D\sigma} = 0.7071 s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

The standard errors of average deviations and differences between average deviations are:

$$\sigma_{AD} = \frac{0.6028\sigma}{\sqrt{N-1}} \quad \text{or} \quad \frac{0.6028s}{\sqrt{N}}$$
$$\sigma_{D_{AD}} = \sqrt{\sigma_{AD_1}^2 + \sigma_{AD_2}^2} \quad \text{or} \quad 0.6028 s_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Any probable error may be obtained by multiplying the corresponding standard error by 0.6745.

The application of the standard error to correlation statistics is discussed on pages 405 to 420.

T—THE NUMBER OF STANDARD ERRORS

For brevity, a deviation from any statistical measure in terms of its standard error is called *T*. The deviations are always considered positive. This definition may be written diagrammatically as follows:

---

[6] **The quartiles Q_1 and Q_1' are both first quartiles, but in two different samples.**

$$T = \begin{array}{c}\text{A}\\ \text{positive}\\ \text{number}\end{array} = \left[\frac{\begin{array}{c}\text{A deviation of any number}\\ \text{from the arithmetic mean}\\ \text{of a sample—or any other measure}\end{array}}{\begin{array}{c}\text{Standard error of that}\\ \text{sample mean—or of that measure}\end{array}}\right]$$

and algebraically as follows:

$$T = \left|\frac{Ma - X}{\sigma_{Ma}}\right|$$

where X is any value.

For example, if an arithmetic mean is 60 pounds and its standard error is 5 pounds, the deviation of 55 pounds from 60 pounds is 1 standard error and is called $T = 1$ $[(60 - 55) \div 5 = 1]$. The deviation of 50 from 60 pounds is called $T = 2$ $[(60 - 50) \div 5 = 2]$. The deviation of 81 from 60 is called $T = 4.2$ $[(81 - 60) \div 5 = 4.2]$.

The values $T = 1$, $T = 2$, and $T = 4.2$ have definite meanings. When $T = 1$, there are 68.27 chances in 100 that the population mean lies within a range of 5 from the sample mean, 60. When $T = 2$, there are 95.45 chances in 100 that the population mean lies within 10 of the sample mean, 60. When $T = 4.2$, there are more than 99.73 chances in 100 that the population mean lies within 21 of the sample mean, 60. It will be noted that these probabilities are the same as those for 1, 2, or 3 standard errors (table 2).

HYPOTHETICAL MEANS

Up to this point, the population mean has been estimated by considering the probabilities of deviations from the sample mean. Another approach can be made. The population mean can be assumed. For example, when the sample mean is 60, the population mean might be assumed to be 50. In this case, 50 is the hypothetical mean or assumed population mean. Instead of testing the deviation of 50 from 60, one might test the deviation of 60 from 50. It may be more logical to consider a population mean rather than a sample mean as the base of deviations, even though the population mean is hypothetical. Regardless of which is the base for measuring deviation, the size of that deviation and the size of T are no different.

The validity of a hypothetical mean is tested with the use of T. When a hypothetical mean is chosen to represent the population mean, T may be defined as the positive difference between the hypothetical and sample means in terms of the standard error of the mean:

$$T = \binom{\text{Positive}}{\text{number}} = \frac{\text{Difference between hypothetical and sample means}}{\text{Standard error of the sample mean}}$$

$$T = \left|\frac{Ma - Ma'}{\sigma_{Ma}}\right|$$

where Ma is the sample mean and Ma' is the hypothetical mean.

Certain arbitrary rules have been set up for the interpretation of T. If $T = 2.0$ or more, that is, if the difference between the hypothetical and sample means is 2 or more standard errors, the difference is said to be significant. In terms of probabilities, such significance may be interpreted in two ways: (*a*) The chances are 95.45 or more out of 100 that the mean of any random sample would deviate less from the hypothetical mean than did the mean of the sample studied. (*b*) The chances are only 4.55 or less that such a large deviation from the hypothetical mean could be expected in any random sample. Since the chances of obtaining such a sample from this population are so small, the conclusion might be that the sample is not from this population. However, the assumption was made that the sample was representative of the population. If the difference between sample and hypothetical means is too great to be explained by chance, it follows that the population mean is not so far from the sample mean as was the hypothetical mean.

When $T = 2$, the difference is said to be significant, and the degree of certainty is indicated by the probability 0.9545. When $T = 3$, the difference is said to be very significant, and the probability is 0.9973. Some statisticians prefer to speak of the probability of the population mean being farther away from the sample mean than the hypothetical mean is. If $T = 2$, there is very little chance that the population mean does not fall nearer the sample than the hypothetical mean does. This probability is only 0.0455. When $T = 3$, the corresponding probability is 0.0027.

Up to this point, only the probabilities of integral values of T have been considered. It is just as logical to speak of the values of T for convenient probabilities, such as 4 out of 5 or 9 out of 10. When $T = 1.64$, the chances are 9 out of 10, or 0.90, that the population mean is no farther away from the sample mean than is the hypothetical mean[7] (table 3). Table 3, to the left, gives the probabilities for integral values 1, 2, and 3. These are the same as the probabilities for 1, 2, and 3 standard errors given in table 2. Table 3, center and right, gives the values of T for convenient probabilities such as 0.50, 0.60, and so on. When $T = 1.960$, the corresponding probability is 95 per cent; and $T = 2.576$, 99 per cent.

---

[7] Likewise, there is 1 chance in 10 that it is farther away.

TABLE 3.—VALUE OF T FOR DIFFERENT PROBABILITIES*

Probability	T	Probability	T	Probability	T
0.6827	1.000	0.50	0.674	0.90	1.645
0.9545	2.000	0.60	0.842	0.95	1.960
0.9973	3.000	0.70	1.036	0.98	2.326
		0.80	1.282	0.99	2.576

* From the normal frequency curve.

NULL HYPOTHESIS

The use of T is more important for testing differences between two sample population means than for testing the means themselves. The difference between the population means is assumed to be zero; that is, the hypothetical difference is zero. This is called the *null hypothesis*—the hypothesis that there is no difference. Assume that average weights of samples of hogs in Illinois and Iowa were 239 and 242 pounds, respectively, and the standard error of the difference between these two means was 0.8 pound. The value of T based on the difference would be

$$T = \frac{D_{Ma} - D'_{Ma}}{\sigma_{D_{Ma}}}$$

The actual difference, D_{Ma}, was 3 pounds, and the hypothetical difference, D'_{Ma}, was 0. Then,

$$T = \frac{D_{Ma} - 0}{\sigma_{D_{Ma}}} = \frac{3.0 - 0}{0.8} = 3.75$$

The value of $T = 3.75$ corresponds to a probability of more than 0.9973, which is the probability for 3.0 standard errors (table 3).

If there were no difference between the two *population* means, the average weights of all hogs in Illinois and Iowa, the differences between *sample* means would be less than 3 pounds in more than 99.73 per cent of such samples. The difference would be greater than 3 pounds in a very small proportion of the cases. Since such a large difference could not be expected due to chance alone; there must be some other reason for the difference. Iowa hogs must really be heavier than Illinois hogs. The difference in the samples is large enough to enable one to state conclusively that the hogs in Iowa are larger than the hogs in Illinois.

Differences are generally said to be significant when $T = 1.960$ or more, and very significant when $T = 2.576$ or more.

SMALL SAMPLES

All probabilities stated thus far have been based on the normal frequency curve, which assumes that the number of observations in samples is large. This assumption is rarely fulfilled. The object of sampling is to reduce the number of observations necessary for the desired information. Strictly speaking, no distribution is quite normal because the number of observations is always limited. With a small number of observations, the distribution may depart greatly from normal. As already stated, the value of T for a probability of 95 per cent is 1.960. However, this is true only for samples of 500 or more. When the sample is 20 items, the value of t for a probability of 0.95 is 2.09;

TABLE 4.—VALUES OF t CORRESPONDING TO VARIOUS PROBABILITIES AND DEGREES OF FREEDOM*

Degrees of freedom n	Probability†				Degrees of freedom n	Probability†			
	50 per cent	90 or 10 per cent	95 or 5 per cent	99 or 1 per cent		50 per cent	90 or 10 per cent	95 or 5 per cent	99 or 1 per cent
1	1.000	6.34	12.71	63.66	24	0.685	1.71	2.06	2.80
2	0.816	2.92	4.30	9.92	25	0.684	1.71	2.06	2.79
3	0.765	2.35	3.18	5.84	26	0.684	1.71	2.06	2.78
4	0.741	2.13	2.78	4.60	27	0.684	1.70	2.05	2.77
5	0.727	2.02	2.57	4.03	28	0.683	1.70	2.05	2.76
6	0.718	1.94	2.45	3.71	29	0.683	1.70	2.04	2.76
7	0.711	1.90	2.36	3.50	30	0.683	1.70	2.04	2.75
8	0.706	1.86	2.31	3.36	35	0.682	1.69	2.03	2.72
9	0.703	1.83	2.26	3.25	40	0.681	1.68	2.02	2.71
10	0.700	1.81	2.23	3.17	45	0.680	1.68	2.02	2.69
11	0.697	1.80	2.20	3.11	50	0.679	1.68	2.01	2.68
12	0.695	1.78	2.18	3.06	60	0.678	1.67	2.00	2.66
13	0.694	1.77	2.16	3.01	70	0.678	1.67	2.00	2.65
14	0.692	1.76	2.14	2.98	80	0.677	1.66	1.99	2.64
15	0.691	1.75	2.13	2.95	90	0.677	1.66	1.99	2.63
16	0.690	1.75	2.12	2.92	100	0.677	1.66	1.98	2.63
17	0.689	1.74	2.11	2.90	150	0.676	1.66	1.98	2.61
18	0.688	1.73	2.10	2.88	200	0.675	1.65	1.97	2.60
19	0.688	1.73	2.09	2.86	300	0.675	1.65	1.97	2.59
20	0.687	1.72	2.09	2.84	400	0.675	1.65	1.97	2.59
21	0.686	1.72	2.08	2.83	500	0.674	1.65	1.96	2.59
22	0.686	1.72	2.07	2.82	1,000	0.674	1.65	1.96	2.58
23	0.685	1.71	2.07	2.81	∞	0.674	1.64	1.96	2.58

---

* t distribution was first published in Fisher, R. A., Statistical Methods for Research Workers, p. 137, 1925. The values of t which appear here were taken from Goulden, C. H., Methods of Statistical Analysis, p. 267, 1939.

† The larger probabilities which are given first, 90 per cent, 95 per cent, or 99 per cent, are the probabilities that t *is not* due to chance alone. The smaller probabilities, 10 per cent, 5 per cent, and 1 per cent, are the probabilities that t *is* due to chance alone.

and when 10, $t = 2.26$. This indicates that, with the same probability, 0.95, the size of T increases as the size of the sample decreases.

For large samples, the number of standard errors has been called T, and for small samples, the number of standard errors has been called t.

For any given probability, say 95 per cent, t has a different value for about every size of sample. Values of t for samples ranging in size from 2 to 1,000 observations are given in table 4. The values of t for four probabilities are included. Each of the four is expressed in two ways. The probabilities that t is *not due to chance* are 50, 90, 95, and 99 per cent. The probabilities that *t is due to chance* are 50, 10, 5, and 1 per cent.

The size of the sample is indicated in the first column headed n (table 4). The number of observations, N, in the sample is *practically* the same as n in the table. However, n refers to the *degrees of freedom* rather than to the number of observations.

DEGREES OF FREEDOM

The degrees of freedom are the number of observations which are free to vary after certain restrictions are imposed. In testing the reliability of an arithmetic mean, the degrees of freedom are one less than the number of observations. Since the calculated arithmetic mean is a fixed value, all the observations cannot fluctuate freely and independently of one another. All the observations but one may have any values regardless of the size of their average. However, after all the observations but one are determined, the last one is automatically fixed. It is fixed because it must be such a number that all the numbers will average the arithmetic mean.

In testing single means, medians, quartiles, standard deviations, and the like, the degrees of freedom are always one less than the number of observations, that is $n = N - 1$. The values of t for N and for $N - 1$ are practically the same for 20 or more observations. The differences are small for samples of 15 or even 10.

In testing differences between two arithmetic averages and the like, the degrees of freedom are usually two less than the total number of observations in the two groups.

A summary of the degrees of freedom which should be used for various t tests is given in table 5. The number of degrees of freedom, $N - 1$, for individual means holds for the arithmetic average, medians, quartiles, average deviations, standard deviations, and the like. The number of degrees of freedom for differences between two means holds for differences between arithmetic means, medians, quartiles, and the like.

TABLE 5.—DEGREES OF FREEDOM FOR VARIOUS t TESTS

Measures tested	Degrees freedom, n	Standard error	Measures tested	Degrees freedom, n	Standard error
Individual means	$N-1$	$\frac{\sigma}{\sqrt{N-1}}$ or $\frac{s}{\sqrt{N}}$	Differences between two frequencies in same distribution	$N-1$	$\sqrt{\frac{2(Nf_o-f_o^2)}{N-1}}$
Differences between two means	$(N_1-1)+(N_2-1)$ or N_1+N_2-2	$s_p\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$	Proportions	$N-1$	$\sqrt{\frac{pq}{N-1}}$
Second differences between means	$(N_1-1)+(N_2-1)+(N_3-1)+(N_4-1)$ or $N_1+N_2+N_3+N_4-4$	$s_p\sqrt{\frac{1}{N_1}+\frac{1}{N_2}+\frac{1}{N_3}+\frac{1}{N_4}}$	Differences between two proportions in same distribution	$N-1$	$\sqrt{\frac{2p_oq_o}{N-1}}$
Frequencies	$N-1$	$\sqrt{\frac{Nf-f^2}{N-1}}$	Differences between two proportions in different distributions	$(N_1-1)+(N_2-1)$ or N_1+N_2-2	$\sqrt{p_oq_o\left(\frac{1}{N_1-1}+\frac{1}{N_2-1}\right)}$

USES

Standard errors and t distribution for small samples were developed to test the reliability of statistical measures and their differences. The technique in testing arithmetic averages and differences between two averages can be applied with a little modification to almost any measure whose standard error can be calculated.

Standard errors are more important for small samples than for large samples. Facts are more doubtful when based on small than on large samples. Standard errors are valuable in clearing up doubts but add little to practical certainties.

Since the most important statistical measure is the arithmetic average, most of the application of standard errors is to averages and their differences. The application to differences among averages is probably more important than to the averages themselves. Averages are used to measure central tendency; differences between averages are widely used to show relationships. Analyzing relationships is by far the more important problem of statistics.

The application of standard errors to the tabular method of analyzing relationship is discussed in detail in the next chapter. Since the most important statistical measure in tabular analysis is the arithmetic mean and relationships are shown by differences in these averages, the emphasis is on standard errors of differences between means.

The application of standard errors to correlation analysis is discussed on pages 405 to 420.

CHAPTER 18

APPLICATION OF STANDARD ERRORS TO TABULAR ANALYSIS

The results of tabular analysis are usually in the form of averages. They are simple and easy to understand. For these reasons, too much dependence is frequently placed on them. Most persons forget that, although an average is based on all the observations in a group, it may be greatly different from most of the observations in that group. The reliability of an average depends on (*a*) the number of observations and (*b*) the variability among the observations. An average made up of only a few highly variable items is of little value, whereas an average including a large number of relatively homogeneous items is reliable. Most persons recognize the limitations of averages based on only a few items. However, few persons recognize the effect of variability among the items on the reliability of the average. The standard errors of means, proportions, and other statistical measures take into consideration both the size of and variability within the groups from which the averages are calculated. In testing averages, small standard errors indicate high degrees of reliability. Small standard errors reflect large numbers and/or low variability within groups.

The problem of testing the reliability of the results of tabular analysis was outlined as follows:

1. Reliability of a single mean.
2. Reliability of differences between two means.
3. Reliability of paired differences.
4. Reliability of differences in frequencies and proportions.

Since differences between means are the tools with which tabular analysis shows relationships, the reliability of these differences is by far the most important.

RELIABILITY OF A SINGLE MEAN

Families with less than $1,000 incomes paid 2.45 cents per pound for potatoes during the winter of 1936–1937 in Rochester, New York (table 1). This average, 2.45 cents, was based on only 38 purchases. The problem is to determine the reliability of 2.45 cents as an average price for all people with like incomes in Rochester. Some indication of the reliability of the average can be obtained from its standard error.

In order to calculate the standard error, the standard deviation of the 38 prices must first be determined:

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - (AX)^2} = \sqrt{\frac{233.2972}{38} - (2.45)^2} = \sqrt{0.1369} = 0.37$$

The standard error is

$$\sigma_{M_a} = \frac{\sigma}{\sqrt{N-1}} = \frac{0.37}{\sqrt{38-1}} = \frac{0.37}{6.1} = 0.061$$

TABLE 1.—TESTING RELIABILITY OF AVERAGES IN A ONE-WAY TABLE

RELATION BETWEEN FAMILY INCOME AND THE RETAIL PRICES OF POTATOES, ROCHESTER, NEW YORK, JANUARY-FEBRUARY 1937

Family income	Number of purchases	Average price, cents per pound	Standard deviation, cents per pound	Standard error, cents per pound	Degrees freedom n	95 per cent probability			99 per cent probability		
						Value of t*	t times standard error	Range likely to include average	Value of t*	t times standard error	Range likely to include average
Less than $1,000	38	2.45	0.37	0.061	37	2.03†	0.12¢	2.33–2.57¢	2.71	0.17	2.28–2.62¢
$1,000–1,999.....	138	2.65	0.22	0.019	137	1.98	0.04	2.61–2.69	2.61	0.05	2.60–2.70
2,000–2,999.....	100	2.73	0.25	0.025	99	1.98	0.05	2.68–2.78	2.63	0.07	2.66–2.80
3,000 and over..	78	3.09	0.68	0.077	77	1.99	0.15	2.94–3.24	2.64	0.20	2.89–3.29

* Table 4, page 320.

† For $n = 35$, $t = 2.03$; and for $n = 40$, $t = 2.02$. For $n = 37$, the value of t was estimated to be 2.03. The values of t for 35 to 40 degrees of freedom and 95 per cent probability range from 2.03 to 2.02. Since 37 is nearer 35 than 40, t was assumed to be 2.03.

Usually the differences are so small that such linear interpolations can be used.

A range of 0.06 cent either side of the average, 2.45, would be 2.39 to 2.51. The chances are, roughly, 2 out of 3 that the average retail price for potatoes of all the low-income groups was between about 2.4 and 2.5 cents per pound. This is a preliminary conclusion.

More definite information concerning the reliability of this average may be obtained. The degrees of freedom, n, in the 38 purchases were $N - 1$, or 37. With a probability of 95 per cent, the value of t for $n = 37$ is 2.03 (table 4, page 320). The value of t, 2.03, is merely a number of standard errors. Since 0.061 is 1 standard error, 2.03 standard errors is 0.124 cent ($2.03 \times 0.061 = 0.124$). The range of 0.12 cent either side of the mean is from 2.33 to 2.57 (table 1). The chances are 95 out of 100 that the average retail price paid for potatoes by all families with low incomes was between 2.33 and 2.57 cents per pound.

With a probability of 99 per cent, $t = 2.71$; and 2.71 standard errors were 0.17 cent per pound. The chances are 99 out of 100 that the average price was between 2.28 and 2.62 cents, a range of 0.17 cent on either side of the mean, 2.45 cents per pound (table 1).

The reliability of the other three averages in table 1 was tested in the same manner.

One can be practically certain that the average price for the lowest-income group is within the limits 2.28 and 2.62 cents. However, this range, 0.34 cent, is considerable, about 14 per cent of the average price. This uncertainty in the average is due to the small number of purchases, $N = 38$, and to the variability in the price, $\sigma = 0.37$.

This uncertainty in the average cannot be reduced by changing the variability in the individual prices because the cause of that variability is not known. The average may be made more accurate by increasing the size of the sample. Assume that there were 200 purchases and that the arithmetic average and the standard deviation of the prices were not changed, $Ma = 2.45$ and $\sigma = 0.37$. The standard error would then be

$$\sigma_{Ma} = \frac{0.37}{\sqrt{200 - 1}} = \frac{0.37}{14.1} = 0.026 \text{ cent}$$

instead of 0.061 cent (table 1).

For 99 per cent probability and 199 degrees of freedom, $n = 199$, the value of t is 2.60. Since 0.026 would be 1 standard error, 2.60 standard errors would be 0.068 $(2.60 \times 0.026 = 0.068)$. The chances would then be 99 out of 100 that the average retail price paid by all low-income families would be between 2.38 and 2.52 cents per pound. The range, 0.14 cent, would then be less than one-half that for the 38 actual purchases, 0.34. This demonstrates the importance of the size of a sample in affecting the reliability of its average.

DIFFERENCES BETWEEN TWO MEANS

One-Way Table

In the analysis of a one-way table, there is always the question of whether the difference between two averages is significant. Can it be said that there was a significant[1] difference between the prices paid for potatoes by the medium-income group, 2.65; and the low-income group, 2.45 cents (table 1)? The reliability of the difference between these two average prices, 0.20 cent per pound, can be tested by comparing this difference to the standard error of this difference (table 2). The standard error of the difference may be calculated from the weighted or pooled standard deviation of the prices in the two groups. The weighted standard deviation may be obtained as follows:

---

[1] Research workers speak freely of "significant" and "very significant" differences. In their broad sense, these terms are meaningless because the degree of certainty is not stated. Practically speaking, significant differences are those differences which are great enough so that the chances are 95 out of 100 against their occurrence due to chance alone. The corresponding probability for very significant differences is 99 per cent.

$$\begin{aligned} s_p &= \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2}} \\ &= \sqrt{\frac{38(0.37)^2 + 138(0.22)^2}{38 + 138 - 2}} = \sqrt{\frac{5.2022 + 6.6792}{174}} = \sqrt{0.0683} \\ &= 0.26 \end{aligned}$$

The standard error of the difference is:

$$\begin{aligned} \sigma_{D_{Ma}} &= s_p\sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \\ &= 0.26\sqrt{\frac{1}{38} + \frac{1}{138}} = 0.26\sqrt{0.0263 + 0.0072} \\ &= 0.048 \end{aligned}$$

The calculated value of t may be determined by dividing the difference between the two means by its standard error:

$$\begin{aligned} t &= \frac{Ma_2 - Ma_1}{\sigma_{D_{Ma}}} \\ &= \frac{2.65 - 2.45}{0.048} = \frac{0.20}{0.048} \\ &= 4.2 \end{aligned}$$

TABLE 2.—TESTING THE SIGNIFICANCE OF DIFFERENCES BETWEEN AVERAGES IN A ONE-WAY TABLE IN WHICH THE RELATIONSHIP *IS* CONSISTENT

Relation between Family Income and the Retail Prices of Potatoes, Rochester, New York, January-February 1937

Family income	Number of purchases	Average price, cents per pound	Differences	Pooled standard deviation*	Standard error of difference	Value of t		Significance of difference
						Calculated	Table, 99 per cent	
Less than $1,000..	38	2.45	0.20	0.26	0.048	4.2	2.61	Very significant
$1,000–1,999.....	138	2.65	0.08	0.23	0.030	2.7	2.60	Very significant
2,000–2,999.....	100	2.73	0.36	0.49	0.074	4.9	2.60	Very significant
3,000 and over...	78	3.09						

---

* Standard deviation of the price paid for each class was omitted as it was given in table 1.

Since $t = 4.2$, the difference, 0.20 cent, is greater than 4 standard errors. The significance of such a difference can be interpreted from a table of t distribution (table 4, page 320). The degrees of freedom are the number of purchases in the two groups minus 2:

$$n = N_1 + N_2 - 2 = 38 + 138 - 2 = 174$$

For 174 degrees of freedom and a probability of 99 per cent, the table value of t is approximately 2.61. It is evident that the calculated value of t, 4.2, is greater than the table value of t and corresponds to a probability even higher than 99 per cent. The chances are greater than 99 out of 100 that such a large difference as 0.20 cent could not occur because of chance alone. In the words of the usual research worker, the difference is very significant.[2]

The other differences in table 2 were tested in the same manner and found to be very significant. From the averages, their differences, and tests of significance, the following generalizations were made.

As income increased, the price paid for potatoes also increased. The relationship was consistent. The increase in price for each successive income group was positive and very significant. The significance of the individual differences is indicated by the t test. The significance of the relationship is even greater than indicated by the t test because the differences were always in the same direction.

CONSISTENCY IN RELATIONSHIPS

The application of standard errors to tabular analysis is limited by the inability to compare more than two averages at a time. With standard errors, one can test the differences between the first group and the second, between the second and the third, between the first and the third, or for any other desired combination of two averages. However, the consistency in the relationship of the first to the second and the second to the third is not directly tested. In general, the relationship is significant if the individual differences are significant. If the individual differences are both significant and consistent, the relation-

---

[2] It is quite certain that there is *some* difference between the two groups. The next question is: Of *how much* difference can one be certain (95 per cent probability)? Let A be this unknown difference. Then

$$t = \frac{D_{Ma} - A}{\sigma_{D_{Ma}}};\ D_{Ma} - A = t\sigma_{D_{Ma}};\ A = D_{Ma} - t\sigma_{D_{Ma}}$$

The 95 per cent value of t for $n = 174$ is 1.98.

$$A = 0.20 - 1.98(0.048) = 0.20 - 0.095 = 0.105 \text{ cent.}$$

The chances are 95 out of 100 that the difference between the two income groups is as great as 0.105 cent per pound.

ship is even more significant. Sometimes, the differences between averages are not reliable, and the *relationship is not wholly consistent.* Yet, the *relationship may be significant.* The relationship in table 3 is a case in point. The 354 retail sales of potatoes which were grouped into only 4 classes in table 2 were grouped into 14 income classes in table 3.

TABLE 3.—TESTING THE SIGNIFICANCE OF DIFFERENCES BETWEEN AVERAGES IN A ONE-WAY TABLE IN WHICH THE RELATIONSHIP *IS NOT* CONSISTENT

RELATION BETWEEN FAMILY INCOME AND THE RETAIL PRICES OF POTATOES, ROCHESTER, NEW YORK, JANUARY-FEBRUARY, 1937

Family income	Number of purchases	Average price, cents per pound	Standard deviation of price, cents	Pooled standard deviation, cents	Standard error of difference, cents	Differences between averages, cents	Value of t: Calculated	Value of t: Table, 95 per cent	Value of t: Table, 99 per cent	Significance of difference	Whether relationship is consistent*
Less than $500	7	2.63	0.27								
				0.45	0.227	−0.35	1.5	2.14	2.98	Not significant	No
$ 500– 749	9	2.28	0.51								
				0.41	0.162	+0.18	1.1	2.04	2.76	Not significant	Yes
750– 999	22	2.46	0.34								
				0.29	0.078	+0.11	1.4	2.00	2.66	Not significant	Yes
1,000–1,249	38	2.57	0.25								
				0.31	0.072	+0.23	3.2	2.00	2.65	Very significant	Yes
1,250–1,499	36	2.80	0.36								
				0.28	0.067	−0.19	2.8	2.00	2.65	Very significant	No
1,500–1,749	34	2.61	0.15								
				0.17	0.043	+0.02	0.5	2.00	2.66	Not significant	Yes
1,750–1,999	30	2.63	0.18								
				0.24	0.055	+0.15	2.7	1.99	2.64	Very significant	Yes
2,000–2,249	50	2.78	0.27								
				0.28	0.080	+0.01	0.1	2.00	2.66	Not significant	Yes
2,250–2,499	16	2.79	0.29								
				0.25	0.078	−0.17	2.2	2.02	2.70	Significant	No
2,500–2,749	28	2.62	0.21								
				0.22	0.099	+0.08	0.8	2.04	2.74	Not significant	Yes
2,750–2,999	6	2.70	0.22								
				0.17	0.080	+0.01	0.1	2.07	2.81	Not significant	Yes
3,000–3,999	19	2.71	0.14								
				0.22	0.073	+1.13	15.5	2.03	2.73	Very significant	Yes
4,000–4,999	17	3.84	0.28								
				0.42	0.121	−0.87	7.2	2.00	2.67	Very significant	No
5,000 and more	42	2.97	0.46								

* Assuming that relationship is positive.

The price tended to increase slightly with income, but the relationship was inconsistent (table 3). Families with less than $500 income paid a higher price for potatoes than those in the 3 income classes from $500 to $1,249. Likewise, those with incomes from $1,250 to $1,499 paid more than the families in the 7 income classes from $1,500 to $3,999. Some persons would probably conclude that, with increasing income, families purchased potatoes of higher quality. After the determination of the increase in prices paid by each successively higher income group and the calculation of *t*, there was considerable variation in the degree of significance of these differences. Of the 13 differences, 7 were not significant, and 6 were either significant or very significant (table 3, next to last column).

The consistency of a relationship is shown by whether the differences between successive averages are all in the same direction. If all had been positive, there would have been no question but that families with higher incomes paid a higher price. However, 4 of the differences were negative and 9 were positive. When the price declined and the difference was negative, the relationship was said to be not consistent (table 3, last column).

Of the six differences which proved to be significant, 3 were consistent, and 3 were not consistent. In other words, there were only 3 out of 13 differences which were *both significant and consistent.* Therefore, the 14 averages do not definitely indicate the presence of a relationship. Consequently, it cannot be generalized from table 3 that the retail prices paid for potatoes rise with increasing family income.

When this problem was studied with only four income groups, the relationship was consistent and significant (table 2). In testing tables with standard errors of differences, the student cannot be certain that he has obtained all the possible information until he has reduced the table to a small number of classes. Some tables show no significant relationships until reduced to only two classes. If the relationship does not prove to be significant with two classes, one can be certain that evidence of a relationship is not present.[3]

Two-Way Tables

In a two-way table, there are two relationships that can be studied with standard errors and differences. The complexity of tests of significance depends on the number of averages in the table. The simplest two-way table with two classifications for each independent variable has four subgroup averages.

The effect of small and large production of flaxseed, X_2, and high and low price of cottonseed meal, X_3, on the price of linseed meal, X_1, for 42 years illustrates the simplest form of two-way table (table 4). The lowest price of linseed meal occurred when there was a large crop of flaxseed and a low price of cottonseed meal; and the highest price, when the opposite conditions existed. This would indicate that some relationship existed between the price of linseed meal, X_1, and both the production of flaxseed, X_2, and the price of cottonseed meal, X_3.

The significance of any effects of X_2 and X_3 on X_1 may be tested in

---

[3] This is assuming that a significant curvilinear relationship has not already been found.

TABLE 4.—TESTING SIGNIFICANCE OF DIFFERENCES BETWEEN AVERAGES IN A TWO-WAY TABLE

RELATION OF THE UNITED STATES PRODUCTION* OF FLAXSEED AND THE PURCHASING POWER† OF THE UTICA PRICE OF COTTONSEED MEAL TO THE PURCHASING POWER† OF THE UTICA PRICE OF LINSEED MEAL,§
1897–1938

Production of flaxseed, X_2	Price of cottonseed meal, X_3		Weighted averages
	Low	High	
	Price linseed meal, X_1	*Price linseed meal,* X_1	*Price linseed meal,* X_1
Small	102.1	104.8	*103.5*
Large	91.9	102.0	*97.0*
Weighted averages	*96.8*	*103.3*	*100.0*

* In per cent of trend.

† Index numbers in terms of prices of 30 basic commodities.

§ Bennett, K.R., The Price of Feed, unpublished manuscript, Cornell University, 1940.

the usual manner by obtaining differences in X_1 and using the *t* test (tables 5 and 6). The averages to be compared were arranged in an orderly manner according to the effects of X_2 and X_3 (table 6).

The effect of X_2 on X_1 may be examined by comparing 102.1 and 91.9, which appear in the first column of table 4. In this comparison, the price of cottonseed, X_3, is held constant at a "low" price. Casual

TABLE 5.—SUPPLEMENTARY DATA NECESSARY FOR THE CALCULATION OF *t* IN TABLE 6

Production of flaxseed, X_2	Production of cottonseed meal, X_3					
	Low	High	All	Low	High	All
	Number of years	*Number of years*	*Number of years*	*σ, price linseed meal,* X_1	*σ, price linseed meal,* X_1	*σ, price linseed meal,* X_1
Small	10	10	20	9.0	11.1	10.2
Large	11	11	22	9.6	12.5	12.2
All	21	21	42	10.6	12.0	*

* Not calculated, not used.

examination might lead to the belief that this difference in the price of linseed meal due to the size of the flaxseed crop was significant. With the t test, the difference, −10.2, was found to be significant (tables 4 and 6).

TABLE 6.—DETERMINATION OF t FOR SEVEN DIFFERENCES BETWEEN AVERAGES

ORDERLY ARRANGEMENT OF AVERAGES TO BE COMPARED AND RELATIONSHIPS TO BE STUDIED IN TABLE 4

Averages compared	Relationships	Variable held constant	Differences between average prices	Pooled standard deviation*	Standard error of difference	Value of t		Significance of difference
						Calculated	Table, 95 per cent	
102.1 and 91.9	Effect of X_2 on X_1	X_3 at "low" prices	−10.2	9.8	4.3	2.4	2.1	Significant
104.8 and 102.0		X_3 at "high" prices	− 2.8	12.5	5.5	0.5	2.1	Not significant
103.5 and 97.0		X_3 at "average" prices	− 6.5	11.6	3.6	1.8	2.0	Almost significant
102.1 and 104.8	Effect of X_3 on X_1	X_2 at "small" crops	+ 2.7	10.7	4.8	0.6	2.1	Not significant
91.9 and 102.0		X_2 at "large" crops	+10.1	11.7	5.0	2.0	2.1	Almost significant
96.8 and 103.3		X_2 at "average" crops	+ 6.5	11.6	3.6	1.8	2.0	Almost significant
91.9 and 104.8	Effect of X_2 and X_3 on X_1	None	+12.9	10.9	4.8	2.7	2.1	Significant

* The pooled standard deviation for the four subgroups was calculated as follows:

$$s_p = \sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_3\sigma_3^2+N_4\sigma_4^2}{N_1+N_2+N_3+N_4-4}}$$

$$= \sqrt{\frac{10(9.0)^2+10(11.1)^2+11(9.6)^2+11(12.5)^2}{10+10+11+11-4}}$$

$$= 11.21$$

Reading down the second column, one can observe the effect of X_2 on X_1, when X_3, the price of cottonseed, is held constant at a "high" level (table 4). The difference, −2.8 (102.0 − 104.8 = −2.8), was not significant (table 6).

Reading down the third column, one can observe the average effect of X_2 on X_1 for both high and low prices of cottonseed meal, X_3. The difference, −6.5, between 97.0 and 103.5 was almost significant (table 6).

In common parlance, the effect of the size of the flaxseed crop on the price of linseed meal was significant only when the price of cottonseed meal was low.

Likewise, the effect of X_3 on X_1 may be examined by comparing the averages 102.1 with 104.8 when X_2 is small; 91.9 with 102.0 when X_2 is large; and 96.8 with 103.3 for all values of X_2. None of the differences were significant although two were almost significant (table 6).

The combined effects of X_2 and X_3 on X_1 are measured by the difference between 91.9 when X_2 was large and X_3 was low, and 104.8 when X_2 was small and X_3 was high.[4] The difference, 12.9, was significant (table 6). Although there is some doubt as to the significance of the effect of either X_2 or X_3 on X_1, their combined effects were significant.

The production of flaxseed and the price of cottonseed meal may have been jointly related to the price of linseed meal. The price of linseed meal was lower when production of flaxseed was large than when it was small. This was true regardless of whether the price of cottonseed meal was low or high. However, when the price of cottonseed meal was low, the effect of large over small crops of flaxseed was −10.2, whereas, when cottonseed meal was high, the corresponding effect was −2.8 (tables 4 and 6). A joint relationship may be said to exist because the effect of the flaxseed crop on the price of linseed meal was different when the price of cottonseed meal was low from that when it was high.

Since a joint relationship exists when there is a difference between the differences, −10.2 and −2.8, the joint relationship may be tested by testing this second difference. The second difference was 7.4 [−2.8 − (−10.2) = 7.4]. The standard error of a second difference is as follows:

$$\begin{aligned}\sigma_{D_1-D_2} &= s_p\sqrt{\frac{1}{N_1}+\frac{1}{N_2}+\frac{1}{N_3}+\frac{1}{N_4}}\\ &= 11.21\sqrt{\frac{1}{10}+\frac{1}{10}+\frac{1}{11}+\frac{1}{11}}\\ &= 11.21\sqrt{0.3818} = 11.21 \times 0.6179\\ &= 6.9\end{aligned}$$

With the null hypothesis, the hypothetical difference is zero, and t is calculated as follows:

$$\begin{aligned}t &= \frac{(D_1 - D_2) - 0}{\sigma_{D_1-D_2}} = \frac{7.4 - 0}{6.9}\\ &= 1.1\end{aligned}$$

---

[4] The combined effects of X_2 and X_3 on X_1 are not measured by the difference between the two averages, 102.1 and 102.0. Since the effects of X_2 were negative and those of X_3 were positive, simultaneous increases in both X_2 and X_3 would tend both to lower and raise the price of linseed meal, X_1. The effects would tend to balance each other.

In this case,

$$
\begin{aligned}
n &= N_1 + N_2 + N_3 + N_4 - 4 \\
&= 10 + 10 + 11 + 11 - 4 = 38
\end{aligned}
$$

or

$$
\begin{aligned}
n &= N - 4 \\
&= 42 - 4 = 38
\end{aligned}
$$

With 38 degrees of freedom, the 95 per cent table value of t was 2.02. Since the calculated value of t, 1.1, was much less than the 95 per cent table value, the second difference was not significant. Since this second difference was not significant, it is dangerous to assume that a joint relationship exists.

The relationships shown by table 4 and tested in table 6 are logical. As the production of flaxseed increases, the price of linseed meal decreases.[5] As the price of cottonseed meal increases, the price of linseed meal also increases.[6] When the production of flaxseed decreases at the same time as the price of cottonseed meal increases, the increase in the price of linseed meal is considerable.[7] However, there is doubt as to the significance of some of these relationships. This does not necessarily mean that the existence of a relationship is disproved. Lack of significance merely indicates that the evidence of a relationship is insufficient. With data for only 42 years, there were only 10 to 11 prices of linseed meal in each group average (table 5).

Diagnosis of a Two-Way Table

Most two-way tables contain more than 4 averages. As many as 25 or 30 averages are not uncommon. When there are 3 classifications for each independent variable, there are 9 *subgroup* averages, exclusive of 6 weighted *group* averages. This is illustrated in table 7. Each of the 9 subgroup averages may be compared with each of the other 8. There are 36 possible comparisons of the 9 subgroup averages alone. To test the significance of every possible difference would involve a great deal of work, much of which would be useless. The problem is to obtain the desired information with a minimum of effort. The student can save much time and labor by detailed examination of tables prior to calculating any standard errors.

In a problem such as the relation of crop yields and size of business to income, it is generally advisable first to examine the differences between the averages for the highest and lowest groups for each independent variable[8] (table 7).

---

[5] Shown by the weighted averages 103.5 and 97.0 (table 4).

[6] Shown by the weighted averages 96.8 and 103.3 (table 4).

[7] Shown by comparison of the diagonal subgroup averages 91.9 and 104.8 (table 4).

TABLE 7.—A TWO-WAY TABLE WHICH MIGHT BE TESTED BY STANDARD ERRORS OF DIFFERENCES

RELATION OF YIELDS AND SIZE OF BUSINESS TO INCOME ON 620 TOBACCO FARMS, VIRGINIA,* 1933

Crop yields, index, X_2	Size of business, total productive man-work units, X_3			
	Less than 350	350–599	600 or more	Average†
	Income, X_1	*Income*, X_1	*Income*, X_1	*Income*, X_1
Less than 85	$ −254	$ −329	$ −564	$ *−356*
85 to 109	−126	−135	−118	*−127*
110 or more	− 83	114	474	*+235*
Average†	*−169*	*−145*	*+ 38*	*− 92*

* Underwood, F. L., Flue-Cured Tobacco Farm Management, Virginia Agricultural Experiment Station, Technical Bulletin 64, p. 222, January 1939.

† Weighted averages.

The difference due to crop yields is very large, +$591 [235 − (−356) = +591]. There is no question but that it should be tested[9] (test 1). If this proves to be significant, the difference due to size of business, $207 [+38 − (−169) = 207], which is not so large, should also be tested[10] (test 2).

These two tests measure the significance of the effect of: (1) X_2 on X_1 without regard to X_3; and (2) X_3 on X_1 without regard to X_2.

The next step is to examine the *subgroups* in table 7 to ascertain whether the relationships are additive or joint. This can be approximated by comparing the differences in the averages for the subgroups with the corresponding differences between groups. For instance, for crop yields, the difference in the *group* averages was +$591; and in the *subgroup* averages, the differences were +$171, +$443, and +$1,038.

---

[8] When the difference between the averages for the highest and lowest groups is significant, the averages for any groups in between usually fall within the range of the highest and lowest. When the range between the averages of the two extreme groups does not include all the averages, the relationship is not consistent and usually not significant.

[9] Assuming that nothing is known of the variability in X_1.

[10] If the difference $591 is not significant, there is no value in testing the difference due to size.

The differences among these differences indicate clearly the presence of a joint relationship. Good crop yields raise income much more on large than on small farms. If the group difference, +$591, was significant,[11] then each of the subgroup differences might be tested (tests 3, 4, and 5). Assume that all the differences are significant; then good crops raise incomes[12] regardless of size of business.

Similarly, for size of business, X_3, the difference between the *group* averages was +$207, and the three differences between *subgroup* averages were −$310, +$8, and +$557. These differences confirm the presence of joint relationship. Large farms return more income than small farms when yields are high, but less when yields are low. Regardless of whether the average difference, +$207, was significant, the two larger subgroup differences, −$310 and +$557, should be tested[13] (tests 6 and 7).

When a joint relationship appears to be present, its significance should be tested. This can be done by testing the difference in the differences, that is, testing the second differences.

The three first differences measuring the effect of size of business for different crop yields were −$310, +$8, and +$557 (reading across the rows of table 7). Since the difference +$8 for medium yields was between the differences −$310 and $557 for poor and good yields, only the latter two differences need be considered. The difference between −$310 and +$557 is +$867 and is called the second difference.[14] Second differences may be tested in the usual manner by calculating their standard errors and using the *t* test (test 8). If the second difference, +$867, proves to be significant, there is little question but that crop yields and size of business were jointly related to the income of Virginia tobacco farms in 1933.

[11] If the difference +$591 was not significant, only the one large subgroup difference, $1,038, should be tested.

[12] Moreover, if a difference as small as +$171 is significant, the difference between the lowest and greatest differences, +$867 (171 − 1,038 = 867), would probably be significant. This would prove the presence of joint relationships.

The second differences may be used to measure the presence of joint relationships (pages 284 and 332).

[13] Obviously, the small difference, +$8, is insignificant.

[14] This second difference was determined by calculating the effects of large over small businesses for poor and good yields, −$310 and +$557, respectively, and then obtaining the difference between these two numbers [+557 −(−310) = +867].

Likewise, this second difference could have been determined by calculating first the effects of good over poor yields when businesses were small and large, +$171 and +$1,038, respectively. The second difference was +$867 (1,038 − 171 = 867). This is the same as the second difference previously calculated.

In testing differences between averages or between differences, it is necessary to calculate:

1. The standard deviation for each group that is compared.
2. The pooled standard deviation for each combination of two groups compared.
3. The standard error of each difference.

All these calculations are laborious, but they must be made before the relatively simple *t* test can be applied. Therefore, before making any detailed calculations, one should do some such "scouting" to determine which differences are worth testing. In the present problem, out of a large number of possible tests, a maximum of eight would verify all the relationships. If some of the eight tests prove differences to be not significant, some of the succeeding tests would not be performed at all.

The general order of testing differences is about as follows:

1. Test the differences in the *group* averages because these group averages (*a*) have the largest number of observations, and (*b*) are likely to be most typical.
2. Test the largest differences first. If the largest differences do not prove to be significant, there would be little point in testing smaller differences.

In general, only differences which appear consistent should be tested. It is ordinarily useless to test differences which do not represent definite relationships.

PAIRED DIFFERENCES

The testing of differences between two means divides itself into two parts.

The first is the testing of the averages of two series which do not contain the same observations but which measure the same phenomenon. This is illustrated by the average price paid for potatoes by different income groups (table 2). The families in the first income group were *not* the same observations as the families in the next higher group. However, the same phenomenon, the price of potatoes, was measured for each group.

The second part is the testing of averages of two series which contain the same observations and which measure the same phenomenon, but under different conditions. This is the problem of paired differences. It is illustrated by the average amounts of hay fed to dairy cows in 20 herds in March and April. The 20 herds were the same observations for both March and April (table 8). The same phenomenon, pounds of hay, was measured for both March and April. The only difference in the averages was the condition, whether March or April.

TABLE 8.—TESTING THE SIGNIFICANCE OF DIFFERENCES BETWEEN AVERAGES FOR PAIRED OBSERVATIONS

Amounts of Hay Fed Dairy Cows per Hundredweight of Milk Produced in 20 Herds During March and April

Herd	Pounds of hay fed during		Difference in pounds	Differences squared
	March	April	March minus April	
1	62	60	+ 2	4
2	58	52	+ 6	36
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
19	75	70	+ 5	25
20	72	73	− 1	1
Total	1,344	1,282	+62	659
Average	67.2	64.1	+ 3.1	32.95

N = Number herds = 20

σ_D = Standard deviation of difference = $\sqrt{32.95 - (3.1)^2}$ = 4.83

n = Degrees freedom = $N - 1$ = 19

σ_{M_a} = Standard error of mean difference = 1.11

t = 2.79 (calculated value)

t = 2.09 (95 per cent table value)

The observations are said to be paired because the hay fed to a herd in March may be compared to the hay fed to the same herd in April.

In testing paired differences, the first step is to calculate the difference for each pair of observations. For herd 1, the amount of hay fed during March was 62 pounds; and during April, 60 pounds. The difference in the hay fed in March over April was +2 pounds (table 8). The average difference, +3.1 pounds, can be obtained in two ways: (*a*) averaging the individual differences (+62 ÷ 20 = +3.1); or (*b*) finding the difference between the averages for March and April (67.2 − 64.1 = +3.1).

The next steps are to calculate the standard deviation in the 20 individual differences, $\sigma = 4.83$, and to obtain the standard error of the mean.[15] The standard error is

$$\sigma_{M_a} = \frac{\sigma}{\sqrt{N-1}} = \frac{4.83}{\sqrt{20-1}} = \frac{4.83}{4.36} = 1.11 \text{ pounds}$$

The next step is to set up the null hypothesis. It is assumed that there is no difference between the amount of hay fed for the two months.

---

[15] This mean, 3.1, is really a mean of differences, a "mean difference," or a "difference between two means." It is called a mean because its reliability is tested with the standard error of the mean.

In other words, the hypothetical mean of the differences is zero. The value of t is

$$t = \frac{Ma - 0}{\sigma_{Ma}} = \frac{3.1 - 0}{1.11} = \frac{3.1}{1.11} = 2.79$$

In the table of t where $n = 19$, and the probability is 99 per cent, $t = 2.86$; and for a probability of 95 per cent, $t = 2.09$ (table 4, page 320). Since the calculated value of t, 2.79, is greater than 2.09 and less than 2.86, the difference[16] between March and April is said to be significant, but not quite very significant.

The student may raise the question as to why the procedures in testing the differences between the averages of *paired* observations and *unpaired* differences are not the same. It is obvious that the method for *paired* differences cannot be used on *unpaired* data. However, the method for *unpaired* data could have been used on *paired* data.

If that procedure had been followed for the hay-consumption problem, the standard error of the difference between the averages for March and April would have been 2.7 pounds;[17] and the calculated value of t, 1.15. The table value of t for a probability of 95 per cent and $n = 38$

---

[16] The research worker is usually most interested in whether the difference is significant, as given above. Some may also wish to know whether the difference is significantly greater than a given amount.

After finding that there is a significant difference between the hay fed in the two months, one could test the hypothesis that there was a difference of, say, 1.0 pound of hay.

$$t = \frac{Ma - 1.0}{\sigma_{Ma}} = \frac{3.1 - 1.0}{1.11} = \frac{2.1}{1.11} = 1.89$$

In the table of t, where $n = 19$ and the probability is 95 per cent, $t = 2.09$. The calculated t, 1.89, is hardly significant. There is not conclusive proof that the difference is as great as 1.0 pound.

How large a difference could one be certain (95 per cent certainty) exists between the two months?

$t = \frac{Ma - A}{\sigma_{Ma}}$; $Ma - A = t\sigma_{Ma}$; and $A = Ma - t\sigma_{Ma}$, where A is that difference.

$A = 3.1 - 2.09(1.11) = 3.1 - 2.32 = 0.78$

One would be 95 per cent certain that the difference is as large as 0.78 pound.

[17] The standard deviations for March and April were 8.4 and 8.1 pounds, respectively. The standard errors were:

March, $\sigma_{Ma} = \frac{8.4}{\sqrt{20 - 1}} = 1.93$ April, $\sigma_{Ma} = \frac{8.1}{\sqrt{20 - 1}} = 1.86$

The standard error of the difference was

$$\sigma_{D_{Ma}} = \sqrt{(1.93)^2 + (1.86)^2} = 2.7$$

The value of t was

$$t = \frac{D_{Ma} - 0}{\sigma_{D_{Ma}}} = \frac{3.1}{2.7} = 1.15$$

is 2.02 (table 4, page 320). The difference would not have appeared significant with this test. However, by the method of *paired* differences, the difference was decidedly significant ($t = 2.79$, table 8).

When the data are *paired*, the method for *unpaired* differences is inefficient; that is, differences may appear to be not significant when they really are significant. In calculating t by the method of *paired* differences, the average difference is compared to a standard error which measures the variability in the individual differences. That variability is entirely due to errors of measurement, chance, and the like. None is due to differences between herds, because each difference compares one herd in March with the *same* herd in April. In calculating t by the method of *unpaired* differences, the same average difference is compared to a standard error of the difference which is ultimately based on the standard deviations of the two groups. These standard deviations measure not only the variability due to errors and chance, but also that due to differences in herds. Hence, the standard error of the difference between two means of *unpaired* items is greater than the standard error of the mean differences between *paired* items. As a result, t is smaller for the unpaired than for the paired method.

In short, when observations can be paired, more definite information is obtained. The method of testing paired differences takes advantage of this, while the method for unpaired differences does not.

RELIABILITY OF DIFFERENCES IN FREQUENCIES AND PROPORTIONS

ONE-WAY FREQUENCY TABLES

Frequency distributions are very common in tabular analysis. These distributions may be in terms of absolute numbers, or percentages, or both. Since distributions in absolute numbers are the most common, attention will first be focused on testing the significance of differences between actual frequencies.[18]

The distribution of 84 farms by the type of lease illustrates this problem (table 9). The 84 farms indicate that crop share leases were by far the most important, more than twice as numerous as stock share leases. The question is whether the indicated difference was significant. The standard error of the difference between two frequencies in the same distribution is given by

$$\sigma_{D_f} = \sqrt{\frac{2(Nf_o - f_o^2)}{N - 1}}$$

---

[18] The standard error of a single class frequency which is not so important as the standard error between two frequencies is given by $\sigma_f = \sqrt{\frac{Nf - f^2}{N - 1}}$, where f is that frequency. The degrees of freedom are $n = N - 1$. The interpretation of σ_f and the value of t derived from it is the same as that for the reliability of a single mean.

where $f_o = \frac{f_1 + f_2}{2}$ and $N =$ all farms. For the crop and stock share leases,

$$f_o = \frac{57 + 23}{2} = \frac{80}{2} = 40$$

and the standard error of the difference between crop and stock share leases was

$$\sigma_{D_f} = \sqrt{\frac{2[(84 \times 40) - (40)^2]}{84 - 1}} = \sqrt{\frac{2 \times 1{,}760}{83}} = \sqrt{42.41} = 6.51$$

With the null hypothesis, the difference is assumed to be zero, and

$$t = \frac{(57 - 23) - 0}{6.51} = \frac{34}{6.51} = 5.2$$

TABLE 9.—TESTING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO FREQUENCIES

DISTRIBUTION OF NON-RELATED TENANT FARMS ACCORDING TO TYPE OF LEASE*

Lease	Number of farms	Difference between frequencies	
			σ_{D_f} = Standard error of difference between first and second frequencies = 6.5
Crop share..	57	34	n = Degrees freedom = $N - 1$ = 83
Stock share.	23		
Cash.......	4		t = 5.2 (calculated value)
Total	84		t = 2.64 (99 per cent table value)

* Schickele, R., and Himmel, J. P., Socio-Economic Phases of Soil Conservation in the Tarkio Creek Area, Iowa Agricultural Experiment Station, Research Bulletin 241, p. 373, October 1938.

Since for a probability of 99 per cent and 83 degrees of freedom, $N - 1$, the table value is $t = 2.64$, the difference is very significant.

Often, the number of observations in a frequency table is expressed as a percentage or proportion (table 10). The difference between the percentages of farms with different types of leases may be tested. The standard error of the difference is

$$\sigma_{D_p} = \sqrt{\frac{2p_oq_o}{N - 1}}$$

where $p_o = \frac{p_1 + p_2}{2}$ and $N =$ all farms. For crops and livestock leases

$$\sigma_{D_p} = \sqrt{\frac{2 \times 0.476 \times 0.524}{(84 - 1)}}$$

where $p_o = \dfrac{0.678 + 0.274}{2} = 0.476.$

$$\sigma_{D_p} = \sqrt{\frac{0.499}{83}} = \sqrt{0.006012}$$
$$= 0.0775$$

The difference between the proportions, 0.404, was divided by the standard error, $\sigma_{D_p} = 0.0775$, to obtain the calculated value of t, 5.2. The difference between the proportion of farms under crop and stock share leases was very significant. The calculated value of t for the proportions, 5.2, was the same as for the frequencies[19] (tables 9 and 10).

TABLE 10.—TESTING DIFFERENCE BETWEEN TWO PROPORTIONS

(After table 9)

Lease	Number of farms	Percentage	Difference between proportions
Crop share....	57	67.8	0.404
Stock share....	23	27.4	
Cash.........	4	4.8	
Total	84	100.0	

σ_{D_p} = Standard error of difference between first and second proportions = 0.0775

n = Degrees freedom = $N - 1$ = 83

t = 5.2 (calculated value)

t = 2.64 (99 per cent table value)

Two-Way Frequency Tables

The differences between the frequencies in a two-way table can be tested in the same manner as in one-way tables. The relation of years of schooling to the residence of farm-reared children is a case in point (table 11). When the standard errors of the differences between group totals are used, certain facts can be observed and verified.

---

[19] In testing the difference between percentages or proportions, the student must note whether the proportions pertain to the same or different frequency distributions. The formula $\sigma_{D_p} = \sqrt{\dfrac{2p_oq_o}{N - 1}}$ applies to proportions in the *same* frequency distributions.

The crop and livestock share leases were two proportions of the same distribution. When the proportions are *not* in the same distribution, the formula is $\sigma_{D_p} = \sqrt{p_oq_o\left(\dfrac{1}{N_1 - 1} + \dfrac{1}{N_2 - 1}\right)}$, where N_1 and N_2 refer to the total observations in two *different* distributions. This formula would be applied in testing the difference between the percentage of farms in Illinois and in Iowa with crop share leases.

TABLE 11.—TESTING THE SIGNIFICANCE OF DIFFERENCES BETWEEN FREQUENCIES IN A TWO-WAY TABLE

RELATION OF EDUCATION TO PRESENT PLACE OF RESIDENCE, ADULT OFFSPRING OF ARKANSAS FARM FAMILIES*

Years in school	Total children	Number of former farm children now living	
		On farms	In towns
10 or less........	358	171	187
Over 10.........	184	65	119
Total	542	236	306

Schooling

$$\sigma_{D_f} = \sqrt{\frac{N^2}{N-1}} = \sqrt{\frac{(542)^2}{541}} = 23.3$$

$$n = N - 1 = 541$$

$$t_{358-184} = \frac{174}{23.3} = 7.5$$

Residence

$$\sigma_{D_f} = \sqrt{\frac{N^2}{N-1}} = \sqrt{\frac{(542)^2}{541}} = 23.3$$

$$n = N - 1 = 541$$

$$t_{306-236} = \frac{70}{23.3} = 3.0 \text{ (calculated value)}$$

$$t = 2.59 \text{ (99 per cent table value)}$$

* Rearranged from McCormick, T. C., Rural Social Organization in South-Central Arkansas, Arkansas Agricultural Experiment Station, Bulletin No. 313, p. 18, December 1934.

The difference between 358 and 184 was very significant ($t_{358-184} = 7.5$), indicating that a majority of the children attended school 10 years or less.

The difference between 236 and 306 was also significant ($t_{306-236} = 3.0$), indicating that the majority of farm children lived in town after they grew up.

The subgroup totals indicate that, regardless of education, a majority of the children lived in town, and, regardless of where the children lived later, the majority attended school 10 years or less.

According to the subtitle of table 11, there was a relation between education and residence. None of the differences tested or mentioned thus far show any such relationship. It is true that children with much schooling tended to live in town rather than on farms, but so did the children with little schooling. It is true that those living on farms tended to have little schooling, but so did those living in towns. The facts observed and verified thus far could have been obtained from two very simple one-way tables of the two sets of group totals. The subgrouping in table 11 contributes very little and does not describe the relationship which is supposed to exist.

The relationship could be shown clearly by changing the subgroup totals or frequencies to percentages (table 12). Of the 358 children with 10 years or less schooling, 48 per cent lived on farms. Of the 184 with over ten years of schooling, only 35 per cent lived on farms. The 48 and 35 percentages are directly comparable, whereas the numbers 171 and 65 are not directly comparable. This is the advantage of proportions.

TABLE 12.—TESTING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN PERCENTAGES IN A TWO-WAY TABLE

(After table 11)

Years in school	Number of children	Percentage of farm children now living		
		On farms	In towns	Total
10 or less	358	48	52	*100*
Over 10	184	35	65	*100*
Total	542	44	56	*100*

Testing difference between the proportions living on farms, 0.48 and 0.35

$$\sigma_{D_p} = \sqrt{p_o q_o\left(\frac{1}{N_1 - 1} + \frac{1}{N_2 - 1}\right)}$$

$$= \sqrt{0.44 \times 0.56\left(\frac{1}{358-1} + \frac{1}{184-1}\right)}$$

$$= \sqrt{0.2464 \times 0.008266}$$

$$= \sqrt{0.002037} = 0.045$$

$$n = N_1 + N_2 - 2$$

$$= 358 + 184 - 2 = 540$$

$$t_{48-35} = \frac{0.13}{0.045} = 2.9 \text{ (calculated value)}$$

$$t = 2.59 \text{ (99 per cent table value)}$$

The difference between 48 and 35 per cent was tested and found to be very significant ($t_{48-35} = 2.9$). The more schooling farm children received, the fewer of them remained on farms.[20] The relationship is significant.

There is a difference between the two approaches to the schooling and residence problem other than between totals and percentages (tables 11 and 12). In the first approach, one frequency was compared with another frequency in the *same distribution.* In the second approach, the percentages were calculated so that the totals for both the "10 or less" and "over 10" groups were 100. The percentage on farms was not compared with the percentage in towns. This comparison, which would have been between two groups in the same classification, would only have shown that there were more in towns than on farms. It would not have shown the relationship between schooling and residence.

The comparison of 35 and 48 per cent involved the difference between two comparable frequencies in *two separate distributions,* "10 or less" and "more than 10" years of school.[21]

---

[20] Stated another way, the more schooling farm children received, the more of them moved to town. The difference between 0.52 and 0.65 is the same as the difference between 0.48 and 0.35 and would be tested in exactly the same manner with the same results.

One-way frequency tables merely describe observations all in the same distribution. Two-way frequency tables also describe two individual distributions, but their primary purpose is to show relationships. In showing relationships, the frequencies must usually be presented as percentages or proportions.

The differences tested in a one-way table are between *actual* or *percentage frequencies* in the *same* distribution. They involve the reliability of a *description.*

The differences tested in a two-way table are between *percentage* frequencies in *different* distributions. They involve the reliability of a *relationship.*

---

[21] This relationship could have been studied equally well by calculating for children on farms and for those in towns the proportions receiving different amounts of education. Then the relationship would have been tested by comparing percentages horizontally in the table, rather than vertically.

CHAPTER 19

THE ANALYSIS OF VARIANCE

The analysis of variance is another method of testing significance. The general objectives of the analysis of variance are the same as for standard errors.[1] However, the analysis of variance may be applied to a wider range of problems. The techniques are different in practice though somewhat similar in theory.

Variance is merely another name for squared standard deviation, σ^2. It is the average of the squared deviations about the arithmetic mean.

$$\text{Variance} = \frac{\Sigma x^2}{N} = \frac{\Sigma(X - AX)^2}{N} = \frac{\Sigma X^2 - \Sigma XAX}{N} = \frac{\Sigma X^2 - (\Sigma X)^2/N}{N}$$

In estimating the universe or population variance from a sample, the sum of the squared deviations is ordinarily divided by the degrees of freedom which are one less than the number of observations, and the formula for variance is:

$$\text{Variance} = \frac{\Sigma x^2}{N - 1} = \frac{\Sigma(X - AX)^2}{N - 1} = \frac{\Sigma X^2 - \Sigma XAX}{N - 1} = \frac{\Sigma X^2 - (\Sigma X)^2/N}{N - 1}$$

SUBDIVISION OF VARIABILITY

The analysis of variance is based on the ability to divide the variability into two or more parts.

The total variance in the cost per hour of horse labor is calculated from the sum of squared deviations for 15 farms and the degrees of freedom. The sums of squared deviations may be calculated from the costs and squares of costs given in table 1.

SUM OF SQUARED DEVIATIONS

For the 15 farms, the sum of the squared deviations[2] in costs was

$$\Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 6{,}136 - \frac{(284)^2}{15}$$
$$= 6{,}136 - 5{,}377 = 759$$

This sum of the squared deviations for all farms may be broken down

---

[1] The development of the standard-error theory preceded analysis of variance.

[2] A method of calculating sums of squares and sums of products with tabulating equipment is given in Appendix B, page 425.

TABLE 1.—CALCULATION OF SUMS OF SQUARED DEVIATIONS AND VARIANCE

COST PER HOUR OF HORSE LABOR ON 15 NEW YORK FARMS, 1937

Farm number	Costs X	X^2
1	24¢	576
2	25	625
3	13	169
4	9	81
5	31	961
6	19	361
7	9	81
8	14	196
9	34	1,156
10	14	196
11	24	576
12	16	256
13	17	289
14	17	289
15	18	324
Total, all	284	6,136
Total, odd	170	4,132
Total, even	114	2,004
Average, all	18.933	
Average, odd	21.250	
Average, even	16.286	

Sum squared deviations $= \Sigma x^2 = \Sigma X^2 - \dfrac{(\Sigma X)^2}{N}$

For all farms, Σx^2 $= 6{,}136 - \dfrac{(284)^2}{15} = 759$

For odd farms $= 4{,}132 - \dfrac{(170)^2}{8} = 519$

For even farms $= 2{,}004 - \dfrac{(114)^2}{7} = 147$

Between odd and even farms $= [8(21.25)^2 + 7(16.286)^2] - \dfrac{(284)^2}{15} = 93$

Variance $= \sigma^2$ $= \dfrac{\Sigma x^2}{N-1}$

For all farms, σ^2 $= \dfrac{759}{15-1} = 54.2$

For odd farms $= \dfrac{519}{8-1} = 74.1$

For even farms $= \dfrac{147}{7-1} = 24.5$

Between odd and even farms $= \dfrac{93}{2-1} = 93$

into several parts. Assume that the 15 farms are divided into two groups on the basis of whether the farm number was odd or even. The sum of squared deviations was

For odd farms, $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N = 4{,}132 - (170)^2/8 = 519$
For even farms, $= 2{,}004 - (114)^2/7 = 147$

The total of these two sums of squares is 666 $(519 + 147 = 666)$. This is not so great as the sum of the squares of the deviations for all the 15 farms, 759. The 759 represents the variability about the average cost for all farms, while the 666 represents the variability within each group about the average for the particular group. The difference between 759 and 666 can be explained by the difference between the averages, 21.250 and 16.286 cents per hour.

The sum of the squared deviations between the two averages is calculated by assuming that each average represents all the items in that group and proceeding in the usual way to obtain the sums of the squared deviations for all 15 farms. There were 8 farms averaging 21.25 and 7 farms averaging 16.286 cents per hour. The sum of the squared deviations is:

$$\text{Sum of squared deviations} = \Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/N = [8(21.25)^2 + 7(16.286)^2] - (284)^2/15 = 3{,}613 + 1{,}857 - 5{,}377 = 93$$

This quantity, 93, is also the difference between 759 and 666. Apparently, the sum of squared deviations about the average for all 15 farms may be divided into three parts:

1. Sum of squared deviations in individual odd farms about the mean of odd farms.
2. Sum of squared deviations in individual even farms about the mean of even farms.
3. Sum of squared deviations in averages for odd and for even farms about the average of all farms.

DEGREES OF FREEDOM

After the sums of squares have been obtained, it is necessary to determine the degrees of freedom. The degrees of freedom are one less than the number of observations.[3] There were 15 farms, and the total degrees of freedom were $N - 1 = 15 - 1 = 14$. Like the total sum of squares, the total degrees of freedom may be divided into parts as follows:

For 8 odd farms, $N - 1 = 7$
For 7 even farms, $N - 1 = 6$

Within the two individual groups, there were 13 degrees of freedom. However, the total was 14. The missing degree of freedom was between the two averages. The average for all 15 farms was considered fixed, 18.9 cents. In relation to the average for all farms, the averages for the odd and even groups may vary. However, there is only one degree of freedom in their variability because, as soon as one group average is determined, the other is also automatically fixed.

The total degrees of freedom in the variability for 15 farms may be divided into these parts:

1. Degrees of freedom in individual odd farms about the average of odd farms.
2. Degrees of freedom in individual even farms about the average of even farms.
3. Degrees of freedom in the averages for odd and for even farms about the average for all farms.

---

[3] With a given arithmetic mean, only one less than the total observations are free to vary. After these are determined, the last observation must be such that the average is as given.

VARIANCE

Variance may be shown diagrammatically as follows:

$$\text{Variance} = \frac{\text{Sum of squared deviations}}{\text{Degrees of freedom}}$$

and algebraically as follows:

$$\sigma^2 = \frac{\Sigma x^2}{N - 1}$$

The total variance for all farms was

$$\sigma^2 = \frac{759}{15 - 1} = 54.2$$

This total variance, 54.2, describes the variability in cost per hour of horse labor. From the three component sums of squares of deviations and the corresponding degrees of freedom, three other variances could be calculated.

1. For odd farms, variance $= \sigma^2 = \frac{519}{7} = 74.1$.
2. For even farms, variance $= \sigma^2 = \frac{147}{6} = 24.5$.
3. Between averages for odd and for even farms, variance $= \sigma^2 = \frac{93}{1} = 93$.

Four different estimates of variance were obtained. When the estimate is based on a part of the variability, it is not likely to be exactly the same as when based on all the variability. The variances ranged from 24.5 to 93. The three variances measure the variability in costs:

1. Among odd farms (74.1).
2. Among even farms (24.5).
3. Between average odd and average even farms (93.0).

The differences in the variances could be due to chance fluctuations in the data or to the factor by which the farms were classified into groups. In this particular case, the differences were probably all due to chance because the odd and even grouping is a chance classification.[4]

RATIO OF TWO VARIANCES

With the analysis-of-variance method of testing significance, one tests the difference between two variances. For example, with the analysis of variance, one might test whether the variance in costs of horse labor on odd farms, 74.1, was significantly greater than the variance on even farms, 24.5. The differences between variances are expressed as ratios, rather than arithmetic differences. Such a ratio might be written as follows:

---

[4] Instead of sorting the costs on the basis of odd- and even-numbered farms, one might have divided the costs of horse labor according to old and young horses, large and small farms, livestock and grain farms, amount of work performed, or the like. The method of calculating variances would have been the same.

$$\text{Variance ratio} = \frac{\sigma^2_{\text{odd}}}{\sigma^2_{\text{even}}} = \frac{74.1}{24.5} = 3.02$$

Whether this ratio is significantly greater than 1.0, which would indicate no difference, might be tested by calculating the standard error of the ratio and applying the *t* test.

The standard error of the ratio of two variances is most accurately stated in terms of logarithms as follows:

$$\sigma_{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right)} = \sqrt{2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

when n_1 and n_2 are the degrees of freedom corresponding to the two variances. The significance of a variance ratio may be determined by comparing the logarithm of the ratio with the standard error of the logarithm of the ratio. The value of *t* is calculated in the usual manner. The null hypothesis is set up as follows:[5]

$$t = \frac{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - 0}{\sigma_{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}}$$

From the table of *t* distribution, the corresponding 95 and 99 per cent values of *t* can be read. By comparing the calculated value of *t* with these table values, the significance in the variance ratio can be determined.[6] In practice, the ratios are not tested in this manner.

---

[5] A hypothetical $\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = 0$ is the same as $\frac{\sigma_1^2}{\sigma_2^2} = 1.0$.

[6] When the variances for odd and even farms are used, the standard error of the logarithm of their ratio is:

$$\sigma_{\log_e\left(\frac{\sigma^2_{\text{odd}}}{\sigma^2_{\text{even}}}\right)} = \sigma_{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right)} = \sqrt{2\left(\frac{1}{7} + \frac{1}{6}\right)} = 0.787$$

and

$$t = \frac{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - 0}{\sigma_{\log_e\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}} = \frac{\log_e\left(\frac{74.1}{24.5}\right) - 0}{0.787} = \frac{\log_e 3.02 - 0}{0.787} = \frac{1.105}{0.787} = 1.40$$

For 13 degrees of freedom and a probability of 95 per cent, the value of *t* is 2.16 (table 4, page 320). The difference between the two variances is not significant; that is, a greater difference than that which existed could have been expected due to chance alone in more than 5 per cent of the samples. In other words, the variability in costs on odd farms was not significantly greater than that on even farms.

TABLE 2.—

5% OR 95% IN LIGHT-FACE TYPE,

n_1 degrees of freedom

Degrees freedom *lesser* variance

n_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41
	98.49	**99.01**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
	34.12	**30.81**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**

---

[*] Snedecor, G. W., Statistical Methods, pp. 184–187, 1940.

VALUES OF F*

1% OR 99% IN BOLD-FACE TYPE

for *greater* variance

14	16	20	24	30	40	50	75	100	200	500	∞	n_2
245	246	248	249	250	251	252	253	253	254	254	254	1
6,142	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**	
19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50	2
99.43	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**	
8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53	3
26.92	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**	
5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	4
14.24	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**	
4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	5
9.77	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**	
3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	6
7.60	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**	
3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23	7
6.35	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**	
3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	8
5.56	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**	
3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	9
5.00	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**	
2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	10
4.60	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**	
2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40	11
4.29	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**	
2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	12
4.05	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**	
2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	13
3.85	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**	
2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	14
3.70	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**	
2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	15
3.56	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**	
2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	16
3.45	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**	
2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	17
3.35	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**	
2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	18
3.27	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**	
2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88	19
3.19	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**	
2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	20
3.13	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**	
2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	21
3.07	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**	
2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	22
3.02	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**	
2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	23
2.97	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**	
2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	24
2.93	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**	
2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	25
2.89	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**	
2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	26
2.86	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**	

Degrees freedom *lesser* variance

TABLE 2.—VALUES

5% OR 95% IN LIGHT-FACE TYPE,

n_1 degrees of freedom

Degrees freedom lesser variance

n_2	1	2	3	4	5	6	7	8	9	10	11	12
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**
1,000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**

OF F (*Continued*)

1% OR 99% IN BOLD-FACE TYPE

for *greater* variance

14	16	20	24	30	40	50	75	100	200	500	∞	n_2
2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	27
2.83	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**	
2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	28
2.80	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**	
2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	29
2.77	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**	
2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	30
2.74	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**	
2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59	32
2.70	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**	
2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	34
2.66	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**	
1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55	36
2.62	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**	
1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53	38
2.59	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**	
1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	40
2.56	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**	
1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49	42
2.54	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**	
1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48	44
2.52	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**	
1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46	46
2.50	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**	
1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45	48
2.48	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**	
1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44	50
2.46	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**	
1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41	55
2.43	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**	
1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39	60
2.40	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**	
1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37	65
2.37	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**	
1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35	70
2.35	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**	
1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32	80
2.32	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**	
1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28	100
2.26	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**	
1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25	125
2.23	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**	
1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22	150
2.20	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**	
1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19	200
2.17	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**	
1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13	400
2.12	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**	
1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08	1,000
2.09	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**	
1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00	∞
2.07	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**	

Degrees freedom *lesser* variance

Since the work involved in making this test is somewhat tedious, convenient tables have been prepared which show 95 and 99 per cent values of the variance ratios for different combinations of degrees of freedom. The tables are in terms of the ratio itself, commonly written $\sigma_1^2/\sigma_2^2 = F$, rather than the logarithms of the ratio. The degrees of freedom at the top of the table always refer to the larger variance, while those at the side refer to the smaller variance (table 2).

The procedure in testing the significance of the difference in variances on odd and even farms with the F test is as follows:

1. Variance ratio is calculated by dividing the larger by the smaller.

$$\text{Variance ratio} = \frac{74.1}{24.5} = 3.02$$

2. The smaller, 24.5, was based on 6 degrees of freedom; and the larger variance, 74.1, on 7 degrees of freedom. The 95 per cent value of F where $n_2 = 6$ and $n_1 = 7$ is 4.21.

3. Since the ratio, 3.02, is less than the 95 per cent table value, 4.21, the difference between the two variances is not significant. The F test gives the same results as the t test with much less work (footnote 6, page 349).

APPLICATIONS OF THE ANALYSIS OF VARIANCE

Difference Between Two Means

The variance based on the averages for odd and even farms, 93, was greater than variances based on the individual odd or individual even farms, 74.1 and 24.5 (table 1). This could be due to chance or to some factor causing the average cost to be greater on odd than on even farms. As the difference between odd and even farms becomes greater, the variance based on the difference between the two averages also becomes greater. Whether the difference between the two averages is significant can be tested by the F test. The variance between averages is compared with the variance within groups on which the averages are based.

The variance within groups is calculated by dividing the sum of the squared deviations within each group by the sum of the degrees of freedom within each group as follows:

	Odd	Even	Sums	Variance within Groups[7]
Sum squared deviations	519 +	147	= 666	= 51.2
Degrees freedom	7 +	6	13	

[7] The variance within groups is a pooled variance calculated from the variability within each of the two groups. This pooled variance is the square of a pooled standard deviation comparable to those calculated in making the t test, page 312.

The ratio of the variance between averages, 93, to the variance within groups is 1.82 (93 ÷ 51.2 = 1.82). For 13 degrees of freedom for the smaller variance and 1 degree for the greater variance, the table value of F for a probability of 95 per cent is 4.67 (table 2, page 350). Since the variance between averages is not significantly larger than the variance within groups, the difference between the average costs is not significant.[8]

The 15 farms in table 1 were also grouped on the basis of hours worked per horse (table 3). Average costs per hour were 23.6 cents for horses working less than 500 hours; and 13.6 cents for over 1,000 hours. The difference in the costs was 10.0 cents per hour. The significance of the difference can be tested by analysis of variance as follows:

1. The total sum of squared deviations is

$$6{,}136 - (284)^2/15 = 759 \text{ (table 1, page 346)}$$

2. The sum of the squared deviations *between* averages is

$$[8(23.625)^2 + 7(13.571)^2] - (284)^2/15 = (4{,}465 + 1{,}289) - 5{,}377$$
$$= 377 \text{ (table 3)}$$

3. The sum of the squared deviations *within* the two groups is the difference between the total and the sum between groups:[9]

$$759 - 377 = 382$$

4. The variance *between* averages is the sum of the squared deviations *between* groups divided by the corresponding degrees of freedom:

$$377 \div 1 = 377$$

5. The variance *within* groups is the sum of squared deviations *within* groups divided by the corresponding degrees of freedom:

$$382 \div 13 = 29.4$$

---

[8] In testing the significance of a variance ratio, the variance within groups is often the basis of comparison. The variance within groups is due to the chance fluctuation of causes not considered. Hence, the variance between groups must be greater than the variance within groups before it can be said that all the difference is not due to chance. How much greater it must be can be read from table 2.

The basis of comparison, which in this case is the variance within groups, is often called *experimental error*.

[9] The sum of squared deviations within the groups could also have been obtained directly by calculating the sum of squared deviations for each group and adding the two. However, it is usually easier to obtain the sum of squared deviations for all observations and then subtract the sum between groups to obtain the sum within groups.

6. The ratio of the variance *between* groups to the variance *within* groups is calculated:

$$377 \div 29.4 = 12.8$$

7. The value of F for a probability of 99 per cent and 13 and 1 degrees of freedom is 9.07. Since 12.8 is greater than 9.07, the difference in the costs per hour was very significant; that is, hours worked had a very significant effect on costs per hour of horse labor.

TABLE 3.—TESTING THE DIFFERENCE BETWEEN TWO MEANS BY ANALYSIS OF VARIANCE

Relation of Hours of Horse Labor to Costs per Hour*

8 farms with less than 500 hours per horse			7 farms with over 1,000 hours per horse		
Farm number*	Costs X	X^2	Farm number*	Costs X	X^2
1	24¢	576	14	17¢	289
12	16	256	7	9	81
9	34	1,156	8	14	196
13	17	289	3	13	169
11	24	576	4	9	81
2	25	625	6	19	361
5	31	961	10	14	196
15	18	324			
Total	189	4,763		95	1,373
Average	23.625			13.571	

Total costs for 15 farms = 284
Total squares for 15 farms = 6,136

Sums of squared deviations

Total $6{,}136 - (284)^2/15 = 759$

Between groups $[8(23.625)^2 + 7(13.571)^2] - (284)^2/15 = 377$

Within groups $759 - 377 = 382$

Degrees of freedom

Total	14
Between groups	1
Within groups	13

Variance

Between groups $\frac{377}{1} = 377$

Within groups $\frac{382}{13} = 29.4$

Variance ratio

F (calculated) $= \frac{\text{Between groups}}{\text{Within groups}} = \frac{377}{29.4} = 12.8$

F (99 per cent table value) = 9.07

* From table 1.

t Test vs. F Test

Both t and F can be used to test the difference between two averages. The results of the tests are exactly the same.[10] However, when there

---

[10] The t test of the difference between costs of horse labor may be calculated as follows:

$$s_p = \sqrt{\frac{298 + 84}{8 + 7 - 2}} = \sqrt{29.38} = 5.42$$

$$\sigma_{D_{Ma}} = 5.42\sqrt{\frac{1}{8} + \frac{1}{7}} = 5.42 \times 0.518 = 2.81$$

$$t = \frac{10.0}{2.81} = 3.56$$

Since the 99 per cent table value of t for 13 degrees of freedom is 3.01, the difference is very significant.

are three or more averages, the analysis of variance with the F test has a decided advantage over standard errors with the t test. With the analysis of variance, the differences among three or more averages can be tested at the same time, whereas with standard errors only two averages can be tested at one time.[11]

ONE-WAY CLASSIFICATION

Simple relationships shown by averages in a one-way classification are easily tested with the analysis of variance (table 4). The relationship between income and the retail prices paid for potatoes was tested by analysis of variance as follows:

1. The total sum of squared deviations is:

$$2{,}741 - (973)^2/354 = 2{,}741 - 2{,}674 = 67$$

2. The sum of squared deviations *among* averages[12] is:

$$\begin{aligned} &38(2.45)^2 + 138(2.65)^2 + 100(2.73)^2 + 78(3.09)^2 - \frac{(973)^2}{354} \\ &= 228.1 \quad + \quad 969.1 \quad + \quad 745.3 \quad + \quad 744.8 \quad - 2{,}674 \\ &= 2{,}687 - 2{,}674 \\ &= 13 \end{aligned}$$

3. The sum of the squared deviations *within* the two groups is the difference between the total and the sum *among* groups.[13]

$$67 - 13 = 54$$

4. The variance *among* averages is the sum of the squared deviations among group averages divided by the corresponding degrees of freedom. There are 3 degrees of freedom, one less than the number of groups, 4,

$$13 \div 3 = 4.3$$

---

[11] To test more than two averages with standard errors, it was necessary to make several separate tests—one for each pair of averages compared (table 3, page 328).

[12] The sum of squared deviations among averages is based on the assumption that the average represents each observation contributing to that average. The calculation is most easily understood when the square of each average is weighted by the number of observations. Furthermore, this method makes use of the average price which usually appears in the table. However, more accuracy is sometimes obtained when the sums, rather than the averages, are used. Instead of multiplying each average squared by the number of observations, each sum is squared and divided by the number of observations. Using potato-price sums (not given in table 4), the sum of squared deviations among groups would be:

$$\begin{aligned} &\frac{(93)^2}{38} + \frac{(366)^2}{138} + \frac{(273)^2}{100} + \frac{(241)^2}{78} - \frac{(973)^2}{354} \\ &= 227.6 + 970.7 + 745.3 + 744.6 - 2{,}674.4 = 13.8 \end{aligned}$$

[13] The sum within groups may be checked by direct calculation from each group.

5. The variance *within* groups is the sum of squared deviations within groups divided by the degrees of freedom within groups. The degrees of freedom for each group are one less than the number of observations. Hence, the total degrees of freedom within groups are the total number of observations minus the number of groups, or

$$354 - 4 = 350$$

Variance = 54 ÷ 350 = 0.154.

6. The ratio of the variance among groups to the variance within groups is

$$F = 4.3 \div 0.154 = 28$$

7. The table value of F for a probability of 99 per cent and 350 and 3 degrees of freedom is approximately 3.84. Since 28 is much greater than 3.84, the relationship of income to price paid for potatoes is very significant.[14]

The steps in the calculation may be summarized as follows:

	Sum Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	99 Per Cent Table Value F
Among averages	13 (step 2)	3 (step 4)	4.3 (step 4)	28 (step 6)	3.84 (table 2)
Within groups	54 (step 3)	350 (step 5)	0.154 (step 5)	—	—
Total	67 (step 1)	353	—	—	—

TABLE 4.—TESTING THE RELIABILITY OF CONSISTENT RELATIONSHIPS IN A ONE-WAY CLASSIFICATION

Relation between Family Income and Retail Price of Potatoes,* Rochester, New York, January-February, 1937

<table>
<tr><th>Family income X_2</th><th>Number of purchases</th><th>Average price, cents per pound X_1</th><td rowspan="3">Sum of prices, $\Sigma X_1 = 973$
Sum of squares, $\Sigma X_1^2 = 2{,}741$
Sum of squared deviations:
Total = 67
Among averages = 13
Within groups = 54
Variance:
Among averages = 4.3
Within averages = 0.154
Variance ratio, $F = 28$
99 per cent table value, $F = 3.84$</td></tr>
<tr><td>Less than $1,000
1,000–1,999
2,000–2,999
3,000 and over</td><td>38
138
100
78</td><td>2.45
2.65
2.73
3.09</td></tr>
<tr><td>All purchases</td><td>354</td><td>2.75</td></tr>
</table>

* From table 1, page 324.

---

[14] The probability was greater than 99 per cent that such large differences as these could *not* occur due to chance alone. Many workers prefer to state probability in a positive way. The probability was less than 1 per cent that such differences would occur due to chance alone.

The differences shown in table 4 were previously tested by standard errors.[15] In the *t* test, each average was compared with the one preceding and the one following. Three tests were made. With analysis of variance, the tests of significance are made by one series of operations. Standard errors tested only one difference at once, whereas the analysis of variance tests the whole relationship at once.

Inconsistent Relationships

The relationship of income to price with four group averages was consistent. With every increase in income, the price of potatoes also increased (table 4). The relationship was found to be significant by both

TABLE 5.—SIGNIFICANCE OF AN INCONSISTENT RELATIONSHIP IN A ONE-WAY CLASSIFICATION

RELATION BETWEEN FAMILY INCOME AND RETAIL PRICE OF POTATOES,* ROCHESTER, NEW YORK, JANUARY-FEBRUARY, 1937

Family income X_2	Number of purchases	Average price, cents per pound X_1
Less than $500	7	2.63
500– 749	9	2.28
750– 999	22	2.46
1,000–1,249	38	2.57
1,250–1,499	36	2.80
1,500–1,749	34	2.61
1,750–1,999	30	2.63
$2,000–2,249	50	2.78
2,250–2,499	16	2.79
2,500–2,749	28	2.62
2,750–2,999	6	2.70
3,000–3,999	19	2.71
4,000–4,999	17	3.84
5,000 and more	42	2.97

Sum of prices, $\Sigma X_1 = 973$. Sum of squares, $\Sigma X_1^2 = 2{,}741$.
Number of purchases, $N = 354$.

	Sum squared deviations	Degrees freedom	Variance	Variance ratio F	99 per cent table value F
Among averages	29	13	2.23	20	2.2
Within groups	38	340	0.112	—	—
Total	67	353	—	—	—

* From table 3, page 328.

---

[15] Pages 325 to 327.

the *t* and *F* tests. However, the *t* tests tested only individual differences, and the conclusion that the relationship was significant was based on the consistency of the differences as well as the significance of the differences. It is not necessary that a relationship be consistent in order to prove significant with the analysis-of-variance test.

When the relation of income to potato prices was shown with 14 group averages rather than 4, it did not prove significant with the *t* test.[16] The averages were not consistent with the relationship, and many of the differences were not significant. The analysis-of-variance *F* test on the same classifications gives different results. The work is summarized in table 5. Since the calculated value of *F*, 20, is considerably greater than 99 per cent table value of *F*, 2.2, the relationship proves to be very significant. Apparently the number of groups in the classification did not greatly affect the analysis-of-variance test.

Many relationships based on a large number of averages and about which there is some doubt can be tested more easily and efficiently with the *F* test than with the *t* test. This is shown by a comparison of the *F* test in table 5 and the *t* test in table 3, page 328. With 14 averages, the *t* test indicated that the inconsistent relationship *was not* significant. However, the *F* test shows that this inconsistent relationship *was* significant. By reducing the number of averages from 14 to 4, the relationship appeared significant even with the *t* test. However, the *F* test was about as efficient with the 14 as with the 4 averages.

NON-NUMERICAL VARIABLES

When the independent variable is not numerical, there is often no way of determining whether a relationship is consistent. For example, different types of farms may return different incomes. Vegetable farms might return more than dairy farms; and fruit farms, more than vegetable farms; but there is no way of telling whether this relationship is consistent. For this reason, it would be impossible to test the significance of the whole relationship with standard errors. On the contrary, with analysis of variance, the significance of the whole relationship can be tested equally well for non-numerical independent variables as for numerical.

TWO-WAY CLASSIFICATIONS WITH EQUAL SUBGROUPS

More than One Observation in Each Subgroup

Two-way classifications may be divided into those with (*a*) equal groups, and (*b*) unequal groups. The application of analysis of variance with equal groups is relatively simple. The results of feeding 120 pigs of four different ages three different rations illustrates a two-way relationship (table 6). Each combination of age and ration was represented by an equal number of pigs, 10. From the average gains, it is evident that: (1) the older pigs gain the faster; and (2) those fed ration *C* gain faster than those fed *B*, and both gained faster than those fed *A*.

---

[16] Pages 327 to 329.

TABLE 6.—TESTING THE RELATIONSHIPS IN A TWO-WAY CLASSIFICATION WITH EQUAL GROUPS

GAINS MADE BY 120 PIGS OF FOUR DIFFERENT AGES FED THREE DIFFERENT RATIONS. TEN PIGS IN EACH GROUP

Age in weeks	Ration			Average	Number of pigs	Total gain, pounds	
	A	*B*	*C*				
	Pounds gain per pig	*Pounds gain per pig*	*Pounds gain per pig*	*Pounds gain per pig*			For 120 pigs: $\Sigma X = 4{,}106$ $\Sigma X^2 = 148{,}746$
9.......	28.6	27.7	32.8	29.7	30	891	
10.......	31.4	31.5	31.0	31.3	30	939	
11.......	33.2	36.1	40.1	36.5	30	1,094	
12.......	37.3	38.4	42.5	39.4	30	1,182	
Average..	32.6	33.4	36.6	34.2	—	—	
Number pigs	40	40	40	—	120	—	
Total gain, pounds..	1,305	1,337	1,464	—	—	4,106	

The relationships of age and ration to gain may be tested as follows:

1. Total sum of squared deviations,

$$148{,}746 - \frac{(4{,}106)^2}{120} = 148{,}746 - 140{,}494 = 8{,}252$$

2. Sum of squared deviations *among* 12 groups, 10 pigs each,[17]

$$\frac{286^2 + 277^2 + 328^2 + 314^2 + 315^2 + 310^2 + 332^2 + 361^2 + 401^2 + 373^2 + 384^2 + 425^2}{10}$$

$$- \frac{4{,}106^2}{120} = 142{,}859 - 140{,}494 = 2{,}365$$

---

[17] The sum of squares between groups was previously calculated by weighting the squared group averages by the number of observations. Multiplying the squared average by the number of observations is exactly the same as dividing the squared total by the number of observations. The totals for each lot of 10 pigs may easily be determined by multiplying the average gains by 10 (table 6).

3. Sum of squared deviations *within* groups,[18]

$$8{,}252 - 2{,}365 = 5{,}887$$

4. Sum of squared deviations due to ages of pigs.

The effect of age is measured by the variability among the averages for the age groups. The sum of the squared deviations due to age may be calculated from the four averages or totals as follows:

$$30\left(\frac{891}{30}\right)^2 + 30\left(\frac{939}{30}\right)^2 + 30\left(\frac{1{,}094}{30}\right)^2 + 30\left(\frac{1{,}182}{30}\right)^2 - 120\left(\frac{4{,}106}{120}\right)^2$$

$$= \frac{891^2 + 939^2 + 1{,}094^2 + 1{,}182^2}{30} - \frac{4{,}106^2}{120}$$

$$= \frac{793{,}881 + 881{,}721 + 1{,}196{,}836 + 1{,}397{,}124}{30} - \frac{16{,}859{,}236}{120}$$

$$= 142{,}319 = 140{,}494 = 1{,}825$$

5. Sum of squared deviations due to ration fed,

$$\frac{1{,}305^2 + 1{,}337^2 + 1{,}464^2}{40} - \frac{4{,}106^2}{120} = 140{,}847 - 140{,}494 = 353$$

6. Sum of squared deviations due to discrepance.

The sum of squared deviations due to age and to rations, 1,825 and 353, respectively, are really parts of the sum of squared deviations among the 12 groups of pigs, 2,365. Ages and ration account for 2,178 out of 2,365. The remainder, 187, is the discrepance (2,365 – 1,825 – 353 = 187).

Discrepance is a measure of the variability among the 12 group averages which is not explained by the 4 age- and the 3 ration-group averages. The average gain made by any one of the 12 groups of 10 pigs could be estimated from the age- and ration-group averages. For example, 9-week pigs gained 4.5 pounds less than average (34.2 – 29.7 = 4.5). Pigs fed ration *A* gained 1.6 pounds less than average (34.2 – 32.6 = 1.6). According to these average relationships, 9-week pigs fed ration *A* should have gained 28.1 pounds each (34.2 – 4.5 – 1.6 = 28.1). They actually gained 28.6 pounds. The deviation between the actual and estimated gains was +0.5 pound per pig, and the squared deviation was 0.25. These squared deviations are the discrepance, and they total 187 for the 120 pigs.

7. Degrees of freedom.

Since there are 120 pigs, the total degrees of freedom are 119 (120 – 1 = 119).

---

[18] Could be obtained independently from the 12 individual groups.

Among the 12 groups, there are 11 degrees of freedom (12 − 1 = 11).

Within the 12 groups, there are 108 degrees of freedom (120 − 12 = 108) or (119 − 11 = 108).

The 11 degrees of freedom among groups may be subdivided. Since there are 4 age groups, the degrees of freedom due to age are 3 (4 − 1 = 3). Likewise, 2 degrees of freedom are due to ration (3 − 1 = 2).

The degrees of freedom for discrepance are obtained as follows:

$$\begin{aligned} & \text{Degrees of freedom} \\ \text{Degrees freedom for discrepance} &= \begin{pmatrix}\text{Among}\\ \text{groups}\end{pmatrix} - \begin{pmatrix}\text{Due to}\\ \text{ration}\end{pmatrix} - \begin{pmatrix}\text{Due to}\\ \text{age}\end{pmatrix} \\ &= 11 - 2 - 3 \\ &= 6 \end{aligned}$$

or

$$\begin{aligned} & \text{Degrees of freedom} \\ \text{Degrees freedom for discrepance} &= \begin{pmatrix}\text{Due to}\\ \text{ration}\end{pmatrix} \times \begin{pmatrix}\text{Due to}\\ \text{age}\end{pmatrix} \\ &= 2 \times 3 \\ &= 6 \end{aligned}$$

8. Variance.

The variances can easily be calculated from the sums of squared deviations and degrees of freedom. The calculation of variances among ages, among rations, in the discrepance, and within groups is summarized below.

9. Basis of comparison.

In calculating the variance ratios, the variances in age, in ration, and in discrepance are compared with variance due to chance fluctuation of causes not considered. In this case, the basis of comparison is the variance within groups, 54.5.

10. The variance ratios, F, were calculated and compared with the corresponding 95 and 99 per cent table values of F as follows:

	Sum Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	95 Per Cent Value F	99 Per Cent Value F
Among ages	1,825	3	608.3	11.16	2.69	3.96
Among rations	353	2	176.5	3.24	3.08	4.80
Discrepance	187	6	31.2	0.57	—	—
Within groups	5,887	108	54.5	←*Basis of comparison* (Experimental error)		
Total	8,252	119				

11. Conclusions.

Comparison of the calculated values of F with the 95 and 99 per cent table values of F indicates that:

(*a*) The differences due to ages were very significant.

(*b*) The differences due to ration were significant.

(*c*) The discrepance was not significant.[19]

Of the three conclusions, the second, (*b*), is the most important. The real object of this two-way classification is to analyze the effect of ration. The age factor was considered in order to isolate the variability due to age. This reduced the variance within groups, making the variance among rations appear more significant. This ability to isolate and discard parts of the variability in phenomena makes the analysis of variance a more efficient test than standard errors.

Difference Between Two Subgroups. The analysis-of-variance test indicated significant differences due to ration. However, it gave no information concerning the differences between any two rations. It is reasonably certain that pigs fed ration C gained more than those fed either A or B (compare 36.6 with 32.6 and 33.4). But there is little difference between pigs fed A and B (compare 32.6 and 33.4 pounds). The significance of the difference between rations A and B can be tested with a little additional work.

Since, for rations A and B, $\Sigma X = 1{,}305 + 1{,}337 = 2{,}642$, the sum of squared deviations between rations A and B is

$$\frac{1{,}305^2 + 1{,}337^2}{40} - \frac{2{,}642^2}{80} = 87{,}265 - 87{,}252 = 13$$

Since, between the A and B ration groups, there is 1 degree of freedom, the variance is 13 $(13 \div 1 = 13)$. The ratio of this variance, 13, to the variance within groups, 54.5, is 0.24 $(13 \div 54.5 = 0.24)$. Since F is less than 1.0, the difference between rations A and B is not significant.

Since the differences among rations A, B, and C were found to be significant, but between rations A and B not significant, the difference between C and (A and B) must be significant. This deduction can be verified. The total sum of squared deviations among all rations, 353, may be divided into that (1) between A and B, and (2) between C and the average of A and B, each with 1 degree of freedom. If the sum between (A and B) is 13, the sum between C and (A and B) is 340 ($353 - 13 = 340$). The corresponding variance is likewise 340 ($340 \div 1 = 340$). The ratio of 340 to the "experimental error" (variance within groups), 54.5, is $F = 6.24$. Since the corresponding 95 per cent value of F is 3.93, the difference between ration C and the other two is significant.

---

[19] When the discrepance is not significant, it is sometimes combined with the variance within groups to form the basis of comparison. Following this procedure, the variance for comparison $= \frac{187 + 5{,}887}{6 + 108} = \frac{6{,}074}{114} = 53.3$ with 114 degrees of freedom.

The justification for this procedure is that, when discrepance is not significant, it is probably due to chance, as is the variability within groups. A significant discrepance indicates the presence of a joint relationship, discussed on pages 374–7.

One Observation in Each Subgroup

There are occasionally two-way classifications with only one observation in each subgroup. Most of these are the results of planned experiments with no replications.

In the field of prices, seasonal variation is an example of a two-way classification with one observation in each subgroup. The two classifications are usually years and months.[20] The student is primarily interested in the differences among months. The purpose of the year classification is to eliminate from the problem the variability due to year which is often greater than the variability due to the months and tends to obscure it.

TABLE 7.—TWO CLASSIFICATIONS WITH ONE OBSERVATION IN EACH GROUP

SEASONAL VARIATION IN THE PERCENTAGE THAT THE PRICE OF WHEAT BRAN WAS OF THE PRICE OF CORN MEAL*

Year	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Total
1929...........	80	81	82	82	79	80	89	85	75	67	72	65	937
1930...........	65	68	69	69	69	80	83	71	62	57	60	64	817
1931...........	66	79	82	81	82	88	95	86	74	70	73	76	952
1932...........	78	79	81	86	91	97	89	76	74	82	87	79	999
1933...........	82	83	79	83	90	100	99	92	98	89	86	80	1,061
1934...........	77	79	81	81	81	81	77	81	71	65	65	62	901
1935...........	59	69	76	74	75	77	77	71	74	80	67	63	862
1936...........	72	79	83	87	83	85	83	73	59	64	56	55	879
1937...........	76	98	94	97	99	99	90	90	83	76	76	75	1,053
Total	655	715	727	740	749	787	782	725	670	650	642	619	8,461
Average	72.8	79.4	80.8	82.2	83.2	87.4	86.9	80.6	74.4	72.2	71.3	68.8	78.3

For the 108 months: $\Sigma X^2 = 674{,}367$; $\Sigma X = 8{,}461$.

---

* Bennett, K. R., The Price of Feed, unpublished manuscript, Cornell University, 1940.

[20] Any one year would appear 12 times, once for each month. Likewise, any one month would appear in every year (9 times in table 7). However, a given month in a given year, such as October 1929, is never replicated in any month of any year. There is one and only one October 1929.

The average price of wheat bran in terms of corn meal varied from a low of 68.8 in September to a high of 87.4 in March (table 7). The averages for these and other months for 1929–1937 indicate the presence of seasonal variation. Whether these seasonal fluctuations are purely chance or really due to the season may be tested by the analysis of variance as follows:

1. Total sum of squared deviations,

$$674{,}367 - \frac{8{,}461^2}{108} = 674{,}367 - 662{,}857 = 11{,}510$$

2. Sum of squared deviations among years,

$$\frac{937^2+817^2+952^2+999^2+1{,}061^2+901^2+862^2+879^2+1{,}053^2}{12} - \frac{8{,}461^2}{108} = 667{,}648 - 662{,}857 = 4{,}791$$

3. Sum of squared deviations among months,

$$\frac{655^2+715^2+727^2+740^2+749^2+787^2+782^2+725^2+670^2+650^2+642^2+619^2}{9} - \frac{8{,}461^2}{108}$$
$$= 666{,}736 - 662{,}857 = 3{,}879$$

4. Sum of squared deviations for discrepance,

$$11{,}510 - 4{,}791 - 3{,}879 = 2{,}840$$

5. Degrees of freedom.
   Total 107 (108 − 1 = 107)
   Among yearly averages 8 (9 − 1 = 8)
   Among monthly averages 11 (12 − 1 = 11)
   Discrepance 88 (107 − 8 − 11 = 88) or (8 × 11 = 88)
6. Variance.

The variances are all calculated by dividing the sums of squared deviations by their corresponding degrees of freedom (given below).

7. Basis of comparison.

In calculating the variance ratio, the variance among years and among months is compared with the variance in the discrepance. When there is only one observation in each subgroup, there is no variance within groups, and discrepance must be used as a basis of comparison. It is assumed that variance in the discrepance measures the variability due to chance fluctuations of causes not considered. In this case, the basis of comparison is 32.3.

8. The variance ratios, *F*, were calculated and compared with the 99 per cent values of *F* as follows:

	Sum Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	99 Per Cent Value of F
Among years	4,791	8	598.9	18.5	2.72
Among months	3,879	11	352.6	10.9	2.46
Discrepance	2,840	88	32.3	←*Basis of comparison*	
Total	11,510	107			

9. Conclusions.

(*a*) The differences among years were very significant.

(*b*) The differences among months were very significant.

Since the primary object was to test the seasonal variation, the second conclusion is the more important.[21] Since the differences among months are very significant, it can be said that the tendency for the price of bran to be high in the spring and low in the fall is very significant.

Difference Between Two Subgroups. Based on all the months, the index of seasonal variation is very significant, but this tells nothing about differences between individual months or groups of months. For instance, the differences between November and March and/or between February and March may be tested as follows:

Sum of squared deviations between November and March.

$$\frac{715^2 + 787^2}{9} - \frac{(715 + 787)^2}{18} = 125{,}622 - 125{,}334 = 288$$

Sum of squared deviations between February and March.

$$\frac{749^2 + 787^2}{9} - \frac{(749 + 787)^2}{18} = 131{,}152 - 131{,}072 = 80.$$

	Sum Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	95 Per Cent Value F
Between November and March	288	1	288	8.92	3.95
Between February and March	80	1	80	2.48	3.95
Discrepance	2,840	88	32.3←*Basis of comparison*		

The increase in the index from November to March was very significant. The indexes for the three intervening months indicate that the increases were consistent.

According to the test, the increase from February to March was not significant. However, since the increase from February to March is a part of the consistent and significant increase from November to March, the February-March increase is probably more significant than the test indicated.

---

[21] The value of F between years need not have been calculated.

OTHER APPLICATIONS

There are innumerable applications of the analysis of variance. The analysis of variance can be applied to three-way and higher-order tables. It may be applied to either additive or joint multiple relationships. With some modifications, the analysis of variance can be used where the numbers of observations in subgroups are neither equal nor proportional.

Some applications of the analysis of variance to different types of multiple relationships in three-way tables with unequal subgroups are given on pages 381 to 386.

✓ RELATIVE MERITS OF *t* AND *F* TESTS

A standard error always tests one difference at a time, whereas the analysis of variance tests a whole series of differences at once. With analysis of variance, a relationship can be tested in one set of operations, but with standard errors, each individual difference must be tested separately. Therefore, with standard errors, the conclusion as to the relationship depends to a considerable extent on the ability of the student to combine several definite values of *t*. On the contrary, with analysis of variance, usually little or no judgment is required to interpret a single variance ratio.

When relationships are tested with standard errors, the number of groups must frequently be reduced to two, three, or four, if the relationship is to prove significant. With the analysis of variance, the relationships also appear more significant as the number of groups decreases. However, analysis of variance will show the relationship to be significant with a larger number of groups than could be used with standard errors.[22]

In effect, the standard error compares the variability between two averages with all the remaining variability in the data. The analysis of variance compares the variability between two averages with only part of the remaining variability. For the standard error, the basis of comparison is all the variability within the two groups compared. For the analysis of variance, the basis of comparison is only the part which cannot be isolated and attributed to some other factor. For these reasons, the F test is more efficient than the t test even when only two groups are to be compared.[23]

---

[22] This is illustrated in tables 1, 2, and 3, pages 324 to 328, and tables 4 and 5, pages 358 and 359.

The order of analysis in the F test is the reverse of that in the t test. With the F test, the whole relationship is tested first, and the detailed examination of its parts follows only if desired, whereas, with the t test, the detailed examination must be made before any conclusions can be reached concerning the whole relationship. Generally, the worker is more interested in the whole relationship than in the examination of its parts.

Usually the amount of work involved in the two tests is not very different.

Both the t and F tests are applicable to testing averages, variability, and correlation. However, the t test can be applied to frequencies and percentages, while the F test cannot.

Where both the t and F tests can be used, the F test is more versatile, more efficient, and gives more complete information.

The only reason that the t test has been and probably is the more widely used is that the F test has been developed more recently.

---

[23] This is not true when the differences can be paired. Paired differences can be tested with either t or F with equal efficiency.

CHAPTER 20

APPLICATION OF ANALYSIS OF VARIANCE TO TABULAR ANALYSIS

ONE-WAY TABLES

The testing of a relationship in a one-way table has already been discussed.[1] The relationship was tested by comparing the variance among the groups with the variance within groups. The method is the same for any number of groups. The sizes of the groups may be equal or unequal. The relationship may be linear or curvilinear. It is not necessary to reduce the number of groups and increase their size in order to test the significance of the relationship. However, a relationship which does not prove significant with many groups will sometimes appear significant with only two groups. The opposite may be true when the relationship is curvilinear.

TWO-WAY TABLES

The analysis of variance lends itself readily to testing two-way tables with equal numbers of observations in subgroups. This application has already been discussed.[2] The example used, the relation of age and ration to gains made by pigs, is typical of experimental results in the biological sciences. It is in these fields that the analysis of variance has, thus far, had its greatest application. The application of the analysis of variance to problems in the social sciences is complicated by the inability to plan experiments. In the problem of pig gains, there were exactly 10 pigs in each lot or subgroup because it was planned that way. In such a problem as the relation of land class and type of road to the value of farms, the numbers within subgroups would not be equal or proportionate because good and poor land and good and poor roads were not placed where they are by research workers. In recent years, statisticians have developed methods of approximating the analysis of variance from two-way and higher-order tables with unequal subgroups.

UNEQUAL SUBGROUPS

The relation of labor efficiency and milk markets to incomes illustrates a two-way table with unequal and disproportionate subgroups

---

[1] Pages 357 to 360, tables 4 and 5.

[2] Pages 360 to 365, table 6.

(table 1). The usual method of presenting the relationship is shown in the first half of table 1. The number of farms for each subgroup, which is given in italics in the second half of the table, indicates that the size of subgroups varied greatly.

TABLE 1.—TWO-WAY TABLE WITH UNEQUAL SUBGROUPS

RELATION OF LABOR EFFICIENCY* AND MILK MARKET TO INCOMES ON 707 FARMS, TOMPKINS COUNTY, NEW YORK, 1927

Labor efficiency, X_3	Milk market, X_2					
	None	Butterfat	Milk	None	Butterfat	Milk
	Income, X_1	*Income,* X_1	*Income,* X_1	*Number farms*	*Number farms*	*Number farms*
Less than 140	\$−136	\$−149	\$− 78	*65*	*82*	*54*
140–200	+ 45	+172	+182	*38*	*77*	*131*
More than 200	+503	+477	+763	*20*	*72*	*168*

* Productive man-work units per man.

In calculating the sums of squared deviations and variances among the groups, each of the nine averages is for the moment considered as one observation. This simplifies the procedure. The nine averages and

TABLE 2.—TESTING RELATIONSHIPS IN A TWO-WAY TABLE WITH UNEQUAL SUBGROUPS

RELATION OF LABOR EFFICIENCY AND MILK MARKET TO INCOMES ON 707 FARMS, TOMPKINS COUNTY, NEW YORK,* 1927

Milk market X_2	Labor efficiency, X_3		Number farms	Income,† X_1		Totals of average incomes, X_1, for	
	Range	Average		Total	Average	Milk market X_2	Labor efficiency X_3
None	Less than 140	102	65	\$− 8,870	\$− 136		\$− 363
	140–200	170	38	+ 1,710	+ 45	\$+ 412	
	More than 200	281	20	+ 10,060	+ 503		
Butterfat	Less than 140	102	82	− 12,210	− 149		+ 399
	140–200	170	77	+ 13,240	+ 172	+ 500	
	More than 200	281	72	+ 34,320	+ 477		
Milk	Less than 140	102	54	− 4,230	− 78		+1,743
	140–200	170	131	+ 23,870	+ 182	+ 867	
	More than 200	281	168	+128,170	+ 763		
Totals	—	—	707	+186,060	+1,779	+1,779	+1,779

* From table 1. † $\Sigma X_1^2 = 604,938,400$.

the totals of averages for different milk markets and labor efficiency were arranged in an orderly fashion to facilitate the analysis of variance in table 2.

The first part of the procedure, which considers each average as one observation, is as follows:

1. Sum of squared deviations among the nine averages:

$$(-136)^2 + 45^2 + 503^2 + (-149)^2 + 172^2 + 477^2 + (-78)^2 + 182^2 + 763^2 - \frac{(1{,}779)^2}{9} = 1{,}174{,}221 - 351{,}649 = 822{,}572$$

2. Sum of squared deviations due to milk market:

$$\frac{412^2 + 500^2 + 867^2}{3} - \frac{(1{,}779)^2}{9} = 390{,}478 - 351{,}649 = 38{,}829$$

3. Sum of squared deviations due to labor efficiency:

$$\frac{(-363)^2 + 399^2 + 1{,}743^2}{3} - \frac{(1{,}779)^2}{9} = 1{,}109{,}673 - 351{,}649 = 758{,}024$$

4. Sum of squares due to discrepance:

$$822{,}572 - 38{,}829 - 758{,}024 = 25{,}719$$

5. Basis of comparison:

The basis of comparison for the variances due to milk market, labor efficiency, and discrepance is the variability within the nine subgroups. The sum of squared deviations within subgroups must be calculated by the same method used for equal subgroups.

(*a*) The total sum of squared deviations for 707 farms,

$$604{,}938{,}400 - \frac{186{,}060^2}{707} = 604{,}938{,}400 - 48{,}965{,}097 = 555{,}973{,}303$$

(*b*) The sum of squared deviations among nine groups is

$$\frac{(-8{,}870)^2}{65} + \frac{1{,}710^2}{38} + \frac{10{,}060^2}{20} + \frac{(-12{,}210)^2}{82} + \frac{13{,}240^2}{77} + \frac{34{,}320^2}{72} + \frac{(-4{,}230)^2}{54} + \frac{23{,}870^2}{131} + \frac{128{,}170^2}{168} - \frac{186{,}060^2}{707}$$
$$= 129{,}265{,}256 - 48{,}965{,}097 = 80{,}300{,}159$$

(*c*) Sum of squared deviations within the nine groups,

$$555{,}973{,}303 - 80{,}300{,}159 = 475{,}673{,}144$$

This sum of squared deviations within subgroups is not comparable to the sums of squared deviations calculated under steps 2, 3, and 4, which were obtained by considering each subgroup average as one observation. The sum of squared deviations within the nine subgroups, 475,673,144, was calculated from all 707 observations. It can be made comparable to the sums in 2, 3, and 4 by dividing by the average number of farms in each subgroup. The average used in this case is the harmonic mean.

(*d*) Harmonic mean, number of observations in subgroups,

$$Mh = \frac{9}{\frac{1}{65} + \frac{1}{38} + \frac{1}{20} + \frac{1}{82} + \frac{1}{77} + \frac{1}{72} + \frac{1}{54} + \frac{1}{131} + \frac{1}{168}}$$

$$= \frac{9}{0.162876} = 55.257$$

(*e*) Sum of squares within subgroups divided by the harmonic mean of the number of farms within subgroups,

$$475{,}673{,}144 \div 55.257 = 8{,}608{,}378$$

When this quantity, 8,608,378, is used as a basis of comparison, the calculation of variances and variance ratios proceeds in the same manner as with equal subgroups.

6. Degrees freedom:

Total = 706 (707 − 1 = 706).
Among subgroups 8 (9 − 1).
  Due to milk market 2 (3 − 1).
  Due to efficiency 2 (3 − 1).
  Due to discrepance 4 (8 − 2 − 2 = 4) or (2 × 2 = 4).
Within subgroups 698 (707 − 9 = 698) or (706 − 8 = 698).

The degrees of freedom are no different with unequal or with equal subgroups.

7. Variances:

The variances due to milk market, labor efficiency, discrepance, and the basis of comparison are calculated by dividing the sums of squared deviations by the corresponding degrees of freedom.

8. Variance ratio:

The variance ratios, F, and the corresponding 95 and 99 per cent values of F are summarized as follows:

Source of Variation	Sums Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	Value of F *95 per cent*	Value of F *99 per cent*
Due to milk market	38,829	2	19,415	1.57	3.01	4.64
Due to labor efficiency	758,024	2	379,012	30.73	3.01	4.64
Due to discrepance	25,719	4	6,430	0.52	—	—
Within subgroups	8,608,378	698	12,333	←——*Basis of comparison*		

9. Conclusions:

The relation of milk market to income was not significant. As would be expected, farms selling milk returned more income than those selling butterfat or no dairy products, but the differences were not large enough to be considered significant.[3]

The relation of labor efficiency to income was very significant.

The discrepance was not significant.

ADDITIVE RELATIONSHIPS

Additive relationships were measured by average incomes for milk market groups and those for efficiency groups. The increase from \$412 to \$867 measures the average or additive relationship of market to income. The difference in these totals of averages was tested with sums of squared deviations and variances *due to milk market.* The variance ratio $F = 1.57$ indicates that the *additive* relationship of market to income was not significant. The variance ratio $F = 30.73$ indicates that the *additive* relationship of efficiency to income was very significant.[4]

JOINT RELATIONSHIPS

The relationship of soils and fertilizer to crop yields is shown in a two-way table (table 3). It is clear that fertilizer has little effect on yield when applied to soils A. On the other hand, the application of fertilizer to soil B has a decided effect on yields. The effect of fertilizer on yields is related to the soil. In other words, the relation of soil and fertilizer to

TABLE 3.—TWO-WAY TABLE SHOWING JOINT RELATIONSHIP

RELATION OF SOIL AND FERTILIZER TO CROP YIELDS ON 134 NEW YORK FARMS

Value per acre of fertilizer applied, X_3	Soil type, X_2	
	A	B
	Crop index, X_1	*Crop index,* X_1
Less than \$2.00	101.1	86.6
\$2.00 to \$3.99	102.6	96.9
\$4.00 or more	103.5	105.3

---

[3] The student may wonder why a difference upwards of \$150 [(+867 − 412) ÷ 3 = 152, table 2] would not be significant with such a large number of farms. However, it must be remembered that variability in incomes is very great.

[4] Joint relationships are indicated by the discrepance which in this case was not significant.

yields is joint. Joint relationships can be tested by the analysis of variance.

The average crop indexes were given in table 3 as they frequently appear in publication. The material was rearranged, and additional information was included in table 4 to facilitate the analysis of variance.

TABLE 4.—TESTING JOINT RELATIONSHIP IN A TWO-WAY TABLE

RELATION OF SOIL AND FERTILIZER TO INDEX OF CROP YIELDS ON 134 NEW YORK FARMS

Soil type, X_2	Value per acre fertilizer applied, X_3	Number farms	Crop yields,* index, X_1		Totals of average crop yields, X_1, for	
			Total	Average	Soil	Fertilizer
A	Less than $2.00	19	1,921	101.1		187.7
	$2.00 to $3.99	26	2,668	102.6	307.2	
	$4.00 or more	12	1,242	103.5		199.5
B	Less than $2.00	24	2,078	86.6		208.8
	$2.00 to $3.99	30	2,907	96.9	288.8	
	$4.00 or more	23	2,422	105.3		

* For the 134 farms, $\Sigma X_1 = 13,238$ and $\Sigma X_1^2 = 1,330,059$.

In calculating the sums of squared deviations due to soils, fertilizer, and discrepance, the average indexes of yields for each subgroup were used as six single observations.

1. Sum of squared deviations among the six averages:

$$101.1^2 + 102.6^2 + 103.5^2 + 86.6^2 + 96.9^2 + 105.3^2 - \frac{596.0^2}{6}$$

$$= 59,437.5 - 59,202.7 = 234.8$$

2. Sum of squared deviations due to soil type:

$$\frac{307.2^2 + 288.8^2}{3} - \frac{596.0^2}{6} = 59,259.1 - 59,202.7 = 56.4$$

3. Sum of squared deviations due to fertilizer application:

$$\frac{187.7^2 + 199.5^2 + 208.8^2}{2} - \frac{596.0^2}{6} = 59,314.5 - 59,202.7 = 111.8$$

4. Sum of squares due to discrepance:

$$234.8 - 111.8 - 56.4 = 66.6$$

5. Basis of comparison:

The basis of comparison for the variances due to soil, fertilizer, and discrepance is the variability within the six subgroups. The sum of the squared deviations for basis of comparison is that within groups divided by the harmonic mean number of observations in subgroups.

(*a*) Total sum of squared deviations:

$$1{,}330{,}059 - \frac{13{,}238^2}{134} = 1{,}330{,}059 - 1{,}307{,}796 = 22{,}263$$

(*b*) Sum of squared deviations among subgroups:

$$\frac{1{,}921^2}{19} + \frac{2{,}668^2}{26} + \frac{1{,}242^2}{12} + \frac{2{,}078^2}{24} + \frac{2{,}907^2}{30} + \frac{2{,}422^2}{23} - \frac{13{,}238^2}{134}$$
$$= 194{,}223 + 273{,}778 + 128{,}547 + 179{,}920 + 281{,}688 + 255{,}047 - 1{,}307{,}796 = 5{,}407$$

(*c*) Sum of squared deviations within subgroups:

$$22{,}263 - 5{,}407 = 16{,}856$$

(*d*) Harmonic mean number of farms in each subgroup:

$$Mh = \frac{6}{\frac{1}{19} + \frac{1}{26} + \frac{1}{12} + \frac{1}{24} + \frac{1}{30} + \frac{1}{23}} = \frac{6}{0.2929} = 20.48$$

(*e*) Sum of squared deviations within subgroups divided by harmonic mean of the number of farms within subgroups:

$$16{,}856 \div 20.48 = 823.0$$

6. Variance:

The variances due to soil, fertilizer, discrepance, and the basis of comparison are calculated by dividing the sum of squared deviations by the corresponding degrees of freedom.

7. Variance ratio:

The variance ratios, F, and the corresponding 99 per cent values of F are summarized as follows:

SOURCE OF VARIATION	SUM SQUARED DEVIATIONS	DEGREES FREEDOM	VARIANCE	VARIANCE RATIO, F	99 PER CENT VALUE F
Soil, additive effect	56.4	1	56.4	8.77	6.84
Fertilizer, additive effect	111.8	2	55.9	8.69	4.78
Discrepance, interaction, or joint effect	66.6	2	33.3	5.18	4.78
Within subgroups	823.0	128	6.43←*Basis of comparison*[5]		

---

[5] Sometimes called experimental error.

8. Conclusions:

The additive effect of soil was very significant (compare 8.77 with 6.84).

The additive effect of fertilizer was very significant (compare 8.69 with 4.78).

The discrepance, the joint effect of soil and fertilizer, was also very significant (compare 5.18 with 4.78). When the discrepance is not significant, it is usually ignored as chance variability. When the discrepance is significant, as here, it is interpreted as the interaction of the two factors. This interaction is merely the joint effect of soil and fertilizer on yields. Whereas variances due to soil and fertilizer measure the additive relationships, the variance due to discrepance or interaction measures the joint relationship of soil and fertilizer on yields.

The procedure in testing tables with joint relationships is no different from that in testing those with only additive relationships. The only difference is in the conclusions drawn from the discrepance. If there is no joint relationship, the discrepance will not be significant; that is, it will be due to chance alone and not to joint effects. If a relationship is joint, that fact will appear in the form of a significant discrepance.

Curvilinear Relationships

In analyzing the relationship between milk market, crop yield, and income to test curvilinear relationships, the data were arranged in an

TABLE 5.—TESTING A CURVILINEAR RELATIONSHIP

Relation of Milk Market and Crop Yields to Incomes, 707 Farms, Tompkins County, New York, 1927

Milk market X_2	Crop yields, X_3		Number farms	Income*, X_1		Total of average incomes, X_1, for	
	Range	Average		Total	Average	Milk market X_2	Crop yields X_3
None	Less than 90	69.0	56	\$− 6,990	\$− 125		\$+ 87
	90–109	99.2	33	+ 2,260	+ 68	\$+ 167	
	110 or more	126.7	34	+ 7,630	+ 224		
Butterfat	Less than 90	69.0	92	+ 1,770	+ 19		
	90–109	99.2	70	+ 9,010	+ 129	+ 504	+ 521
	110 or more	126.7	69	+ 24,570	+ 356		
Milk	Less than 90	69.0	118	+ 22,830	+ 193		+1,344
	90–109	99.2	124	+ 40,170	+ 324	+1,281	
	110 or more	126.7	111	+ 84,810	+ 764		
Total	—	—	707	+186,060	+1,952	+1,952	+1,952

* $\Sigma X_1^2 = 604,938,400$.

orderly manner to facilitate calculations (table 5). Regardless of whether the relationships are linear or curvilinear, the analysis of variance at first proceeds in the usual manner:

1. Sum of squared deviations among the subgroup averages:

$$(-125)^2 + 68^2 + 224^2 + 19^2 + 129^2 + 356^2 + 193^2 + 324^2 + 764^2 - \frac{1{,}952^2}{9} = 940{,}084 - 423{,}367 = 516{,}717$$

2. Sum of squared deviations due to milk market:

$$\frac{167^2 + 504^2 + 1{,}281^2}{3} - \frac{1{,}952^2}{9} = 640{,}955 - 423{,}367 = 217{,}588$$

3. Sum of squared deviations due to crop yields:

$$\frac{87^2 + 521^2 + 1{,}344^2}{3} - \frac{1{,}952^2}{9} = 695{,}115 - 423{,}367 = 271{,}748$$

4. Sum of squared deviations due to discrepance:

$$516{,}717 - 217{,}588 - 271{,}748 = 27{,}381$$

5. Basis of comparison:

When the sums of squares and sums for all 707 farms are used, a comparable sum of squared deviations within groups is calculated as follows:

(*a*) Total sum of squared deviations for 707 incomes:

$$604{,}938{,}400 - \frac{186{,}060^2}{707} = 555{,}973{,}303$$

(*b*) Total sum of squared deviations among the nine subgroups:

$$\frac{(-6{,}690)^2}{56} + \frac{2{,}260^2}{33} + \frac{7{,}630^2}{34} + \frac{1{,}770^2}{92} + \frac{9{,}010^2}{70} + \frac{24{,}570^2}{69} + \frac{22{,}830^2}{118} + \frac{40{,}170^2}{124} + \frac{84{,}810^2}{111} - \frac{186{,}060^2}{707} = 45{,}873{,}568$$

(*c*) Sum of squared deviations within the nine subgroups:

$$555{,}973{,}303 - 45{,}873{,}568 = 510{,}099{,}735$$

(*d*) Harmonic mean number of observations in subgroups:

$$Mh = \frac{9}{\frac{1}{56} + \frac{1}{33} + \frac{1}{34} + \frac{1}{92} + \frac{1}{70} + \frac{1}{69} + \frac{1}{118} + \frac{1}{124} + \frac{1}{111}} = \frac{9}{0.142768} = 63.039$$

(*e*) Sum of squared deviations within subgroups divided by harmonic mean number of farms within subgroups:

$$510{,}099{,}735 \div 63.039 = 8{,}091{,}812$$

6. Degrees of freedom, variance, variance ratios, and their significance are summarized as follows:

Source of Variation	Sum Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	Values of F *95 per cent*	Values of F *99 per cent*
Milk market, additive effect	217,588	2	108,794	9.38	3.01	4.64
Crop yields, additive effect	271,748	2	135,874	11.72	3.01	4.64
Discrepance, joint effect	27,381	4	6,845	0.59	—	—
Within subgroups	8,091,812	698	11,593 ←	*Basis of comparison*		

7. Conclusions:

The additive effect of milk market on income was highly significant.[6]

The additive effect of yields on income was highly significant.

The joint effect of milk market and yields on income was not significant.

8. Curvilinearity.

The analysis of variance as applied thus far *tests only* whether a relationship is present. It *does not test* directly the pattern of that relationship.

The relation of yields to income was curvilinear. This is indicated by comparison of changes in yield with changes in income. When average yield increased from 69.0 to 99.2, the total of average income increased from +\$87 to +\$521 (table 5). The average rate of increase was \$4.8 per point increase in crop index $\left(\frac{434}{3} \div 30.2 = 4.79\right)$. When average yields increased from 99.2 to 126.7, the rate of increase in income was \$10 $\left(\frac{823}{3} \div 27.5 = 9.98\right)$. As yields increased, income rose at an increasing rate. In other words, the relationship was curvilinear. Whether the relationship was significantly curvilinear can be tested by analysis of variance as follows:

---

[6] On page 374 it was shown that the effect of milk market on income was not significant. However, in that problem, the effect of labor efficiency was removed. In the present problem, the effect of crop yields was removed, but not the effect of labor efficiency. Since there was an interrelationship between milk market and labor efficiency, the apparent effect of milk market in this problem is partly the effect of labor efficiency. Whether the method of analyzing relationships be tabulation or correlation, the failure to consider all the important independent variables may lead to erroneous conclusions. This is especially true when some of the independent variables are interrelated, as in the above example.

(*a*) A least-squares straight line was fitted to the totals of average incomes, X_1, for different crop yields, X_3, as follows:

X_3	X_1	X_1X_3	X_3^2
69.0	87	6,003.0	4,761.00
99.2	521	51,683.2	9,840.64
126.7	1,344	170,284.8	16,052.89
Σ294.9	1,952	227,971.0	30,654.53

NORMAL EQUATIONS

$$\Sigma X_1 = Na + b_{13}\Sigma X_3$$
$$\Sigma X_1X_3 = a\Sigma X_3 + b_{13}\Sigma X_3^2$$
$$1{,}952.0 = 3a + 294.9b_{13}$$
$$227{,}971.0 = 294.9a + 30{,}654.53b_{13}$$
$$a = -1{,}478.9; \quad b_{13} = +21.664$$

From the equation $X_1 = -1{,}478.9 + 21.664X_3$, the estimated values of X_1 were obtained and compared with the actual as follows:

CROP YIELD X_3	ESTIMATED INCOME X_1'	ACTUAL INCOME X_1
69.0	$16	$ 87
99.2	670	521
126.7	1,266	1,344
Σ294.9	1,952	1,952

On page 378, the sum of squared deviations due to crop yields was 271,748. This sum can be subdivided into two parts—the sum of squares due to a linear relation, and the additional sum of squares due to a departure from linearity. The sum due to the linear relation is calculated from the estimated incomes, X_1', as follows:

$$\frac{16^2 + 670^2 + 1{,}266^2}{3} - \frac{1{,}952^2}{9} = 683{,}971 - 423{,}367 = 260{,}604$$

The sum of the squares due to curvilinearity would be the difference between the total due to yields and that due to the linear relation:

$$271{,}748 - 260{,}604 = 11{,}144$$

The significance of the curvilinearity can be tested by calculating its variance and the variance ratio:

$$\text{Variance} = 11{,}144 \div 1 = 11{,}144$$

and

$$\text{Variance ratio, } F = \frac{11{,}144}{11{,}593} = 0.96$$

Of the two degrees of freedom between crop yield groups, one was allotted to the linear relationship, and the remaining one to curvilinearity. The basis of comparison was the same as that for the problem as a whole.

Since F was less than 1, the curvilinearity was not significant. Although the relationship was somewhat curvilinear, there was no evidence that this departure from linearity was not a random fluctuation.

THREE-WAY TABLES

The three-way table[7] used in chapters 8 and 15 is used here to illustrate tests of significance by the analysis of variance. Size of farm, crop yields, and labor efficiency were related to income (table 6). Incomes increased with larger farms, better crop yields, and higher labor efficiency.

TABLE 6.—THREE-WAY TABULAR ANALYSIS

RELATION OF SIZE, CROP YIELDS, AND LABOR EFFICIENCY TO INCOME, 907 NEW YORK FARMS, 1927

Size of farm, X_2	Crop yields, X_3	Labor efficiency, X_4	
		Low	High
		Income, X_1	*Income*, X_1
Small	poor	\$−119	\$+ 384
Small	good	+101	+ 361
Large	poor	−271	+ 592
Large	good	+232	+1,139

With three-way, four-way, and higher-order tables, the analysis of variance becomes quite detailed, but is not difficult. The detail is due to the large number of relationships involved. In a three-way table, there are three additive and four joint relationships to be tested. The eight average incomes were arranged in one column in order to facilitate calculations (table 7). The analysis of variance for a three-way table involves:

1. Eight averages of the three-way table (column 1).
2. Two totals of averages for each of the three possible one-way classifications (columns 2, 3, and 4).
3. Four totals of averages for each of the three possible two-way classifications (columns 5, 6, and 7).

These averages and totals of averages are all necessary for a convenient system of calculating the analysis of variance. The determination of these averages and totals was as follows:

---

[7] Table 4, page 125, tabular analysis of relationships.
Table 7, page 279, tabulation *vs.* correlation analysis.

TABLE 7.—TESTING RELATIONSHIPS IN A THREE-WAY TABLE

Size farm X_2	Crop yields X_3	Labor efficiency X_4	Average income,* dollars	Totals of average incomes, X_1, in dollars for						Data necessary for sum of squared deviations within groups	
				One-way classification for			Two-way classification for				
				X_2 holding constant X_3 and X_4	X_3 holding constant X_2 and X_4	X_4 holding constant X_2 and X_3	X_2 and X_3 holding constant X_4	X_2 and X_4 holding constant X_3	X_3 and X_4 holding constant X_2	Number farms	Sum incomes X_1 dollars 00 omitted
Small	poor	low	− 119	+ 727	+ 586	− 57	+ 265	− 18	− 390	181	\$− 215
"	"	high	+ 384			+2,476		+ 745.	+ 976	49	+ 188
"	good	low	+ 101		+1,833		+ 462		+ 333	166	+ 167
"	"	high	+ 361						+1,500	46	+ 166
Large	poor	low	− 271	+1,692			+ 321	− 39		45	− 122
"	"	high	+ 592					+1,731		173	+1,025
"	good	low	+ 232				+1,371			68	+ 158
"	"	high	+1,139							179	+2,038
Total			2,419	2,419	2,419	2,419	2,419	2,419	2,419	907	3,405
Column			1	2	3	4	5	6	7	$\Sigma X_1^2 = 108,367$	

* Table 6 rearranged.

In column 2, table 7, the total, +\$727, was the sum of the four averages for small farms; and the total, +\$1,692, the corresponding sum for large farms.

In columns 3 and 4 are the totals of average incomes for poor and good crop yields and for low and high efficiency, respectively. Each total represents the sum of four average incomes indicated by brackets.

In column 5, the total, +\$265, was the sum of the two averages for small farms, X_2, with poor yields, X_3; and the totals +\$462, +\$321, and +\$1,371 were the sums for small X_2 and good X_3, large X_2 and poor X_3, and large X_2 and good X_3, respectively.

Similarly, columns 6 and 7 contain the totals of the two averages for the four combinations of X_2X_4 and X_3X_4, respectively.

The general procedure in testing a three-way table is the same as for two-way tables. The method for unequal subgroups is used. Each of the eight subgroup averages is considered a single observation. The analysis of variance proceeds as follows:

1. The total sum of squared deviations among the eight averages from column 1, table 7, was as follows:

$$(-119)^2 + 384^2 + 101^2 + 361^2 + (-271)^2 + 592^2 + 232^2 + 1{,}139^2 - \frac{2{,}419^2}{8} = 2{,}077{,}189 - 731{,}445.1 = 1{,}345{,}743.9$$

This sum contains the sums of squared deviations for all the additive and joint effects of X_2, X_3, and X_4 on X_1.

2. Additive effects:

VARIABLE	SUMS OF SQUARED DEVIATIONS
X_2 (column 2)	$\frac{727^2 + 1{,}692^2}{4} - \frac{2{,}419^2}{8} = 847{,}848.3 - 731{,}445.1 = 116{,}403.2$
X_3 (column 3)	$\frac{586^2 + 1{,}833^2}{4} - \frac{2{,}419^2}{8} = 925{,}821.3 - 731{,}445.1 = 194{,}376.2$
X_4 (column 4)	$\frac{(-57)^2 + 2{,}476^2}{4} - \frac{2{,}419^2}{8} = 1{,}533{,}456.3 - 731{,}445.1 = 802{,}011.2$

3. The combined additive and joint effects of two variables:

The additive and joint effects of two variables are measured by the sum of squared deviations among the incomes for the four possible combinations of those two variables. For example, the additive and joint effects of X_2 and X_3 are measured by sums of squared deviations among the four group totals in column 5, table 7.

JOINT AND ADDITIVE EFFECTS OF TWO VARIABLES	SUMS OF SQUARED DEVIATIONS
X_2 and X_3 (column 5)	$\frac{265^2+462^2+321^2+1{,}371^2}{2} - \frac{2{,}419^2}{8} = 1{,}133{,}175.5 - 731{,}445.1 = 401{,}730.4$
X_2 and X_4 (column 6)	$\frac{(-18)^2+745^2+(-39)^2+1{,}731^2}{2} - \frac{2{,}419^2}{8} = 1{,}776{,}615.5 - 731{,}445.1 = 1{,}045{,}170.4$
X_3 and X_4 (column 7)	$\frac{(-390)^2+976^2+333^2+1{,}500^2}{2} - \frac{2{,}419^2}{8} = 1{,}732{,}782.5 - 731{,}445.1 = 1{,}001{,}337.4$

4. Joint effects of two variables:

The joint and additive effect of two variables may be subdivided into:

(*a*) Additive effect of first variable.

(*b*) Additive effect of second variable.

(*c*) Joint effect of the two variables.

Since the additive effects of all variables and the combined additive and joint effects of each pair of variables have been determined, the joint effect of two variables can easily be obtained by subtraction.

$$\begin{pmatrix}\text{Joint and additive}\\ \text{effect of } X_2 \text{ and } X_3\end{pmatrix} - \begin{pmatrix}\text{Additive effect}\\ \text{of } X_2\end{pmatrix} - \begin{pmatrix}\text{Additive effect}\\ \text{of } X_3\end{pmatrix} = \begin{pmatrix}\text{Joint effect}\\ \text{of } X_2 \text{ and } X_3\end{pmatrix}$$

VARIABLES	SUMS OF SQUARED DEVIATIONS			
	(joint and additive)	−(additive)	− (additive)	= (joint)
X_2 and X_3	401,730.4	− 116,403.2	− 194,376.2	= 90,951.0
X_2 and X_4	1,045,170.4	− 116,403.2	− 802,011.2	= 126,756.0
X_3 and X_4	1,001,337.4	− 194,376.2	− 802,011.2	= 4,950.0

5. Joint effect of three variables:

The total sum of squared deviations among the eight average incomes was 1,345,743.9 (calculated in 1). This total is composed of the sums of squared deviations for all the possible additive and joint effects of the three independent variables which are:

(*a*) Additive effects of X_2, X_3, and X_4.

(*b*) Two-way joint effects of X_2X_3, X_2X_4, and X_3X_4.

(*c*) Three-way joint effect of $X_2X_3X_4$.

Since the total sum of squared deviations and the sums for additive and two-way joint effects have already been calculated, the sum for the three-way joint effect can be obtained by subtraction.

JOINT EFFECT OF	SUM OF SQUARED DEVIATIONS
X_2, X_3, and X_4	1,345,743.9 −116,403.2 −194,376.2 −802,011.2 −90,951.0 −126,756.0 −4,950.0 = 10,296.3

6. Basis of comparison:

The basis of comparison for testing the various joint and additive effects is the variability within the eight income subgroups.

(*a*) The total sum of squared deviations for the 907 farms was calculated from the sum of their incomes, \$3,405, and the sum of their squared incomes,[8] \$108,367 (table 7).

$$10{,}000\left(108{,}367 - \frac{3{,}405^2}{907}\right) = 955{,}841{,}700$$

---

[8] The work of calculation was simplified by rounding the individual incomes to the nearest \$100. The sum of squared deviations was converted back into terms of dollar incomes by multiplying by 10,000.

(*b*) The sum of squared deviations among the eight income subgroups was calculated from the number of farms and sums of incomes for each group (table 7, right), as follows:

$$10{,}000\left[\frac{(-215)^2}{181}+\frac{188^2}{49}+\frac{167^2}{166}+\frac{166^2}{46}+\frac{(-122)^2}{45}+\frac{1{,}025^2}{173}+\frac{158^2}{68}\right.$$
$$\left.+\frac{2{,}038^2}{179}-\frac{3{,}405^2}{907}\right]=10{,}000(31{,}718.21-12{,}782.83)=189{,}353{,}800$$

(*c*) Sum of squared deviations within the subgroups.

$$955{,}841{,}700-189{,}353{,}800=766{,}487{,}900$$

(*d*) Harmonic mean number of observations in subgroups (table 7, right).

$$\frac{8}{\frac{1}{181}+\frac{1}{49}+\frac{1}{166}+\frac{1}{46}+\frac{1}{45}+\frac{1}{173}+\frac{1}{68}+\frac{1}{179}}=\frac{8}{0.101991}=78.438$$

(*e*) Sum of squared deviations within subgroups divided by harmonic mean number of observations in subgroups.

$$766{,}487{,}900\div 78.438=9{,}771{,}900$$

7. Degrees of freedom:

The total degrees of freedom were 906 (907 − 1 = 906).

The degrees of freedom within subgroups were 899 (907 − 8 = 899).

The degrees of freedom among subgroups were 7 (8 − 1 = 7).

These may be subdivided and allotted to the various additive and joint effects in the same manner as the sums of squared deviations.

(*a*) Additive effect of each variable—1 degree of freedom (2 − 1 = 1).

(*b*) Additive and joint effects of two variables—3 degrees of freedom (4 − 1 = 3).

(*c*) Joint effect of two variables—1 degree of freedom (3 − 1 − 1 = 1).

(*d*) Additive and joint effects of three variables—7 degrees of freedom (8 − 1 = 7).

(*e*) Joint effect of three variables—1 degree of freedom (7 − 1 − 1 − 1 − 1 − 1 − 1 = 1).

The 7 degrees of freedom are finally divided into (*a*) 3 for additive effect, (*c*) 3 for two-way joint effects, and (*e*) 1 for the three-way joint effect.

8. Variances and variance ratios:

The variances due to the different additive and joint effects were calculated from the sums of squared deviations and degrees of freedom.

Using the variance within groups as a basis of comparison, each variance ratio was determined. The calculation and significance of variance ratios were summarized as follows:

Relationship	Sum of Squared Deviations	Degrees Freedom	Variance	Variance Ratio, F	Values of F *95 Per Cent*	Values of F *99 Per Cent*
Additive						
X_2	116,403	1	116,403	10.71	3.85	6.67
X_3	194,376	1	194,376	17.88	Same	Same
X_4	802,011	1	802,011	73.78	for	for
Joint, two variables						
X_2X_3	90,951	1	90,951	8.37	all	all
X_2X_4	126,756	1	126,756	11.66	rela-	rela-
X_3X_4	4,950	1	4,950	0.46	tion-	tion-
Joint, three variables						
$X_2X_3X_4$	10,296	1	10,296	0.95	ships	ships
Basis of comparison	9,771,900	899	10,870	—	—	—

9. Conclusions:

The additive effects of all three factors, size, yields, and efficiency, were very significant.

The joint effect of size, X_2, and yields, X_3, was very significant. The joint effect of size, X_2, and efficiency, X_4, was also very significant. The joint effect of yields, X_3, and efficiency, X_4, was not significant. Likewise, the three-way joint effect of X_2, X_3, and X_4 was not significant.[9]

---

[9] In tables 7 to 14, pages 279 to 288, the data in three-way tables used to illustrate the difference method of analyzing relationships were the same as those used to illustrate the application of analysis of variance. With the difference method, the various additive and joint effects of the three independent variables were calculated from the first, second, and third differences. These effects could also be calculated from the corresponding sums of squared deviations of variances. The procedure would be to obtain the square root of the variance and divide it by 4. Except for signs, the effects obtained by the two methods would check very closely. The variance method does not indicate the direction of the relationship.

Since each variable was divided into only two groups, this example is a special case. The difference method is adaptable to this special case, but could not readily be applied to problems with three or more groups for each variable. However, this example is not a special case in regard to applicability of analysis of variance. Regardless of the number of groups, all additive and joint relationships can be isolated with the analysis of variance.

CHAPTER 21

CHI SQUARE

✓ Chi square is a test of significance for frequencies. Chi square is a calculated measure of the degree to which the frequencies in an actual distribution do not conform to the corresponding frequencies in a theoretical distribution.

GENERAL METHOD

The calculation of chi square, χ^2, may be illustrated by an actual and theoretical distribution of the results of tossing 5 pennies 100 times. The 100 tosses were classified according to the number of heads that came up with the tosses.

Number of Heads	Actual	Theoretical	Deviation	Deviations Squared	Deviations Squared Divided by Theoretical
0	2	3	−1	1	0.3333
1	19	16	+3	9	0.5625
2	30	31	−1	1	0.0322
3	32	31	+1	1	0.0322
4	14	16	−2	4	0.2500
5	3	3	0	0	0.0000
Total	100	100	0	—	1.2102

In 2 of the tosses, all 5 pennies turned up tails. The most numerous numbers of heads were 2 and 3. In the calculation of chi square, χ^2, actual results are compared with the theoretical. In this case, the theoretical frequencies were those that would be expected due to chance alone. The deviations of the actual from the theoretical frequencies were calculated and squared. The squared deviations were divided by the corresponding theoretical frequencies. The sum of these quotients, 1.2102, was χ^2.

Chi square, χ^2, depends directly on how closely the actual and theoretical frequencies agree. When the differences between the two series are large, the deviations squared and χ^2 are large. When the two series of frequencies coincide closely as in the penny-tossing example, χ^2 is small. If the two series coincide exactly, χ^2 would be zero.

TABLE 1.—VALUES OF CHI SQUARE, χ^2, FOR GIVEN PROBABILITIES AND DEGREES OF FREEDOM*

Degrees of freedom, *n*	Probabilities† in percentage							
	99	95	50	30	20	10	5	1
1	0.0002	0.004	0.455	1.074	1.642	2.706	3.841	6.635
2	0.0201	0.103	1.386	2.408	3.219	4.605	5.991	9.210
3	0.115	0.352	2.366	3.665	4.642	6.251	7.815	11.341
4	0.297	0.711	3.357	4.878	5.989	7.779	9.488	13.277
5	0.554	1.145	4.351	6.064	7.289	9.236	11.070	15.086
6	0.872	1.635	5.348	7.231	8.558	10.645	12.592	16.812
7	1.239	2.167	6.346	8.383	9.803	12.017	14.067	18.475
8	1.646	2.733	7.344	9.524	11.030	13.362	15.507	20.090
9	2.088	3.325	8.343	10.656	12.242	14.684	16.919	21.666
10	2.558	3.940	9.342	11.781	13.442	15.987	18.307	23.209
11	3.053	4.575	10.341	12.899	14.631	17.275	19.675	24.725
12	3.571	5.226	11.340	14.011	15.812	18.549	21.026	26.217
13	4.107	5.892	12.340	15.119	16.985	19.812	22.362	27.688
14	4.660	6.571	13.339	16.222	18.151	21.064	23.685	29.141
15	5.229	7.261	14.339	17.322	19.311	22.307	24.996	30.578
16	5.812	7.962	15.338	18.418	20.465	23.542	26.296	32.000
17	6.408	8.672	16.338	19.511	21.615	24.769	27.587	33.409
18	7.015	9.390	17.338	20.601	22.760	25.989	28.869	34.805
19	7.633	10.117	18.338	21.689	23.900	27.204	30.144	36.191
20	8.260	10.851	19.337	22.775	25.038	28.412	31.410	37.566
21	8.897	11.591	20.337	23.858	26.171	29.615	32.671	38.932
22	9.542	12.338	21.337	24.939	27.301	30.813	33.924	40.289
23	10.196	13.091	22.337	26.018	28.429	32.007	35.172	41.638
24	10.856	13.848	23.337	27.096	29.553	33.196	36.415	42.980
25	11.524	14.611	24.337	28.172	30.675	34.382	37.652	44.314
26	12.198	15.379	25.336	29.246	31.795	35.563	38.885	45.642
27	12.879	16.151	26.336	30.319	32.912	36.741	40.113	46.963
28	13.565	16.928	27.336	31.391	34.027	37.916	41.337	48.278
29	14.256	17.708	28.336	32.461	35.139	39.087	42.557	49.588
30	14.953	18.493	29.336	33.530	36.250	40.256	43.773	50.892

---

* Snedecor, G. W., Statistical Methods, p. 163, 1940.

† These are the probabilities that as large a value of χ^2 as that shown would occur as the result of chance alone. The probabilities that as large a χ^2 would *not* occur as the result of chance alone would be given by 1, 5, 50, 70, 80, 90, 95, and 99. For the purpose of testing difference between distributions or between relationships in two- or three-way contingency tables, the latter probabilities are probably the more valuable.

Sometimes the actual frequencies are different from the theoretical owing to chance alone, and sometimes because the theoretical series is not correct. Chi square tests whether the difference could be due to chance. The values of χ^2 for different degrees of freedom and different probabilities have been determined to facilitate the chi-square test (table 1). In table 1, the degrees of freedom at the left are the degrees of freedom between the groups or frequencies. The probabilities at the top of table 1 are the probabilities that such large values of χ^2 as appear in the body of table 1 would occur as the result of chance alone.

For the penny-tossing problem, the degrees of freedom were 5, one less than the number of frequencies, 6. For a probability of 95 per cent, the table value was $\chi^2 = 1.145$. The calculated value of $\chi^2 = 1.210$ was about the same as the 95 per cent table value. A value of χ^2 as large as that calculated, 1.210, would occur because of chance alone in 95 cases out of 100. Since there is nothing unusual about a χ^2 as small as this, the actual distribution is not significantly different from the theoretical. In other words, there is no indication that the theoretical distribution was incorrect.

If the calculated χ^2 had exceeded 11.070, the 5 per cent value, the conclusion would have been the opposite. Such a large value of χ^2 would occur as the result of chance alone in only 5 per cent of the cases. Therefore, the accuracy of the theoretical distribution would have been questioned.

One of the most obvious applications of chi square is testing the goodness of fit of mathematical frequency curves. The observed frequency is compared with the theoretical frequencies read from the fitted curve.

APPLICATION

Chi square has many practical applications, the most important of which are testing differences or relationships. In general, the method of calculating χ^2 is the same for all its applications. The only judgment required from the student is in correctly setting up the problem. With what theoretical distribution should the actual be compared?

For instance, if the type of farm operator was to be studied, some theoretical distribution must be established, and that theoretical distribution depends on the purpose of the problem. The total number of farm operators in Winston County, Alabama, 2,177 (table 2), was distributed as follows:

1. **Full owners, 1,155**
2. **Part owners, 81**
3. **Croppers, 388**
4. **Other tenants, 553**
5. **Managers, 0**

TABLE 2.—DISTRIBUTION OF FARMS ACCORDING TO OPERATOR IN WINSTON COUNTY, ALABAMA, AND THE STATE, 1930*

Operator	State	Winston County
Full owners	75,144	1,155
Part owners	15,228	81
Croppers	65,134	388
Other tenants	101,286	553
Managers	603	0
Total	257,395	2,177

* Fifteenth Census of the United States: 1930, Agriculture, Vol. II, part 2, pp. 978 and 983, 1932.

This distribution may be tested against several theoretical distributions established on the following hypotheses:

1. Out of thin air, Bennett and Pearson say that the distribution is as follows: 50 per cent, 5 per cent, 15 per cent, 25 per cent, and 5 per cent. Since there were 2,177 farms, the theoretical distribution would be: 1,088 full owners, 109 part owners, 327 croppers, 544 other tenants, and 109 managers.

2. On the basis of the normal relation of income to tenure in a given year, the theoretical percentage distribution was 30 per cent, 10 per cent, 20 per cent, 39 per cent, and 1 per cent, which must be used to distribute the total, 2,177.

3. On the basis of the number of farm children in country schools whose fathers were in the various tenure groups, the theoretical percentage distribution *might have been* 41, 5, 26, 28, and 0 per cent.

4. Assuming no differences in the relative importance of these five types of tenure, the theoretical percentage distribution would be 20, 20, 20, 20, and 20 per cent.

5. According to the United States distribution of farms by type of operator, the theoretical distribution would be 46, 11, 12, 30, and 1 per cent.

6. According to the Alabama distribution of farms, the theoretical distribution would be approximately 29, 6, 25, 39, and 1 per cent.

Which one of these six theoretical frequencies is to be used depends on the purpose of the investigation. If the purpose of the investigation is to determine the value of the snap judgment of Bennett and Pearson, use the first hypothesis. If the purpose is to measure the degree of relationship between incomes and tenure, use the second. If the purpose is to test the relation between tenure and number of children per family, use the third. If the purpose is to test the differences among the frequencies, use the fourth. If the purpose is to test whether Winston County differs from the United States as a whole, use the fifth. If the purpose is to compare Winston County with the state of Alabama, use the sixth.

Many more such hypotheses might be formulated which would establish a basis for a theoretical distribution. The excuse for any hypothetical distribution is to test differences or relationships.

TABLE 3.—TESTING DIFFERENCES BETWEEN AN ACTUAL DISTRIBUTION AND A THEORETICAL DISTRIBUTION BASED ON THE POPULATION OF WHICH THE ACTUAL DISTRIBUTION IS A PART*

RELATION OF TENURE IN WINSTON COUNTY, ALABAMA, TO THE TENURE IN THE STATE OF ALABAMA

Operator	State of Alabama		Winston County		Deviations, actual minus theoretical	Deviations squared	Deviations squared divided by theoretical	χ^2 for four degrees of freedom and 1 per cent probability
	Number of farm operators	Per cent of operators	Actual number of farm operators	Theoretical number from Alabama percentages				
Full owner	75,144	29.19	1,155	635	+520	270,400	425.8	
Part owner	15,228	5.92	81	129	− 48	2,304	17.9	
Croppers	65,134	25.31	388	551	−163	26,569	48.2	
Other tenants	101,286	39.35	553	857	−304	92,416	107.8	
Managers	603	0.23	0	5	− 5	25	5.0	
Total	257,395	100.00	2,177	2,177	0	—	$\chi^2 = 604.7$	$\chi^2 = 13.3$

* From table 2.

TESTING WHETHER A FREQUENCY DISTRIBUTION FOR A SAMPLE DIFFERS FROM THAT FOR THE POPULATION

The problem is to test whether tenure in Winston County, the sample, is any different from that in the whole state of Alabama, the population. The farms operated by full owners or tenants for Winston County cannot be compared directly with those for the state because the state is so much larger than any one county. The two series of frequencies can be made comparable as follows: For the state as a whole, the number of farms in each tenure group is expressed as a percentage of the total (table 3). For example, 29.19 per cent of the farms were operated by full owners. The theoretical frequencies for Winston County were obtained by multiplying each of the percentages for the state by the total number of farms in the county. For example, the theoretical number of full owners was 635 (2,177 × 0.2919 = 635), which is directly comparable with the actual number, 1,155. The theoretical frequencies are based on the state totals. In other words, the theory or hypothesis to be tested is that the distribution in the county is the same as that in the state.

To obtain chi square, the deviations of the actual from the theoretical frequencies were calculated and squared. The squares were divided by the theoretical frequencies and the quotients were summed, giving the value of $\chi^2 = 604.7$ (table 3).

Since there were five classes of farms, there were four degrees of freedom. For a probability of 1 per cent, $\chi^2 = 13.277$. Since the calculated $\chi^2 = 604.7$ is much larger, the chances were small that differences between the state and the county were purely random. The hypothesis that the distribution of farms according to tenure in Winston County is no different from that in the state is incorrect. In other words, the difference in tenure was very significant. Tenure in the Republican County, Winston, was not typical of the Democratic state of Alabama.

In the farm-tenure problem, the theoretical frequencies were based on the averages for a large population, the state, of which the county studied was a small part. In other problems, the theoretical frequencies are obtained in a variety of ways depending on the purpose.

TESTING WHETHER FREQUENCIES ARE PROPORTIONAL TO SOME RELATED FACTOR

It was found that the acres per tractor were greater on poor lands (classes II and III, table 4) than on good lands (classes IV and V). In other words, there was a relation between land class and the number of tractors. The question may be raised whether the differences shown in the second column of table 4 are significant. The relation of land class to the number of crop acres per tractor may be tested with chi square.

Chi square cannot be calculated directly from ratios such as acres per tractor. The original data from which the ratio was calculated must be used instead (table 4, center). The actual number of tractors for different land classes is then compared with a theoretical distribution. In this case, the theoretical distribution is to be based on the proportion of the crop area in each land class.

Testing whether acres per tractor are constant on all land classes is the same as testing whether tractors per acre are constant. If tractors per acre are the same on all land classes, then the number of tractors must be proportional to the number of acres. The theoretical distribution was thus based on the number of crop acres.

TABLE 4.—TESTING DIFFERENCES BETWEEN AN ACTUAL DISTRIBUTION AND A THEORETICAL DISTRIBUTION BASED ON SOME RELATED FACTOR

RELATION OF LAND CLASS TO THE NUMBER OF TRACTORS ON FARMS

Relationship		Additional data		Calculation of χ^2				
				Number of tractors				
Land class	Acres per tractor	Number of crop acres	Per cent of crop acres	Actual	Theoretical from per cent of crop acres	Deviations squared, actual less theoretical	Deviations squared divided by theoretical	χ^2 for three degrees freedom and 1 per cent probability
II	97.6	1,561	24.80	16	30	$(-14)^2$	6.533	
III	69.3	2,079	33.04	30	40	$(-10)^2$	2.500	
IV	39.0	1,558	24.76	40	30	$(+10)^2$	3.333	
V	31.3	1,095	17.40	35	21	$(+14)^2$	9.333	
Total	52.0	6,293	100.00	121	121	—	$\chi^2 = 21.699$	$\chi^2 = 11.341$

The percentage of crop acres in each land class was calculated and multiplied by the total number of tractors, 121. For instance, land class II contained 24.80 per cent of the land. The theoretical number of tractors was 24.8 per cent of the total number, 30 (121 × 0.248 = 30). After the theoretical frequencies had been established, χ^2 was calculated in the usual manner and found to be 21.699 (table 4, right). The 1 per cent table value of χ^2 for three degrees freedom was 11.341. Since the calculated value was greater than the table value, the difference between the actual and theoretical distributions was probably not all due to chance. Since the theoretical distribution assumed that the acres per tractor were constant, and since the actual distribution was very significantly different from the theoretical, the acres per tractor were not constant at all; that is, the relation of land classes to number of tractors was very significant.

TESTING DIFFERENCES AMONG FREQUENCIES IN A ONE-WAY TABLE

Chi square can be used to test differences among the individual frequencies in a distribution. The distribution of 84 farm leases among crop share, stock share, and cash leases may be tested. With χ^2, all the differences among crop share, stock share, and cash leases can be tested at once (table 5).

With chi square, an hypothesis is set up that there are no differences

TABLE 5.—TESTING THE DIFFERENCES AMONG THE FREQUENCIES OF ONE DISTRIBUTION

NON-RELATED TENANT FARMS AND TYPE OF LEASE*

Lease	Number of farms		Deviations squared	Deviations squared divided by theoretical	χ^2 for two degrees freedom and 1 per cent probability
	Actual	Theoretical			
Crop share......	57	28	$(+29)^2$	30.04	
Stock share.....	23	28	$(-5)^2$	0.89	
Cash...........	4	28	$(-24)^2$	20.57	
Total	84	84	—	$\chi^2 = 51.50$	$\chi^2 = 9.21$
Average	28	28	—	—	—

* From table 9, page 340.

among the prevalence of the three types of leases.[1] The three theoretical frequencies based on this hypothesis are equal. Since there were 84 farms and three types of leases, the theoretical number of farms for each lease was 28 (84 ÷ 3 = 28) (table 5). Chi square was 51.50, considerably above the 1 per cent value, 9.21. The hypothesis that there is *no* difference is very significantly *not* true; that is, the differences among the numbers of the three leases are very significant.

With standard errors, the difference between two frequencies was tested, whereas with chi square the differences among all three frequencies were tested. Chi square can be employed either to test all the differences at once or to test any individual differences separately. Testing only the difference between the two frequencies 57 and 23, the numbers of farms with crop and stock share leases, the theoretical numbers would be 40 and 40. Chi square would be

$$\chi^2 = \frac{(57 - 40)^2 + (23 - 40)^2}{40} = 14.45$$

Since the 1 per cent table value of χ^2 for one degree of freedom is 6.635, the difference between the numbers of crop and stock share leases was very significant. This corresponds with the results from the *t* test obtained on page 340.

---

[1] This hypothesis is identical to item 4 for the Winston County tenure problem, page 390.

TESTING RELATIONSHIPS SHOWN BY TWO-WAY FREQUENCY TABLES

Relationships shown by a two-way *frequency* table can be tested by χ^2. The relation of education to residence is a two-way frequency to be tested (table 6).

TABLE 6.—TESTING RELATIONSHIP IN A TWO-WAY FREQUENCY TABLE*

RELATION OF EDUCATION TO PRESENT PLACE OF RESIDENCE, NUMBER OF FORMER ARKANSAS FARM CHILDREN NOW LIVING ON FARMS AND IN TOWNS

Relationship tested in a two-way frequency†				Calculations				
Years in school	Present place of residence			Theoretical distribution			Deviations squared divided by theoretical	χ^2 for one degree freedom and 1 per cent probability
				Present place of residence				
	On farms	In towns	Total	On farms	In towns	Total		
	Number former farm children	*Number former farm children*	*Number former farm children*	*Number former farm children*	*Number former farm children*	*Number former farm children*	$\frac{(171-156)^2}{156}=1.442$ $\frac{(65-80)^2}{80}=2.813$ $\frac{(187-202)^2}{202}=1.114$ $\frac{(119-104)^2}{104}=2.163$	
10 or less	171	187	358	156	202	358		
Over 10	65	119	184	80	104	184		
Total	236	306	542	236	306	542	Calculated value $\chi^2=7.532$	$\chi^2=6.635$
Per cent	43.5	56.5	100	43.5	56.5	100		

* Such tables are commonly called "contingency tables," probably because the theoretical frequencies are contingent on totals of actual frequencies for the rows and columns.

† From table 11, page 342.

In testing relationships in two-way frequency tables, the theoretical distribution to which the actual is compared is based on the totals of the table itself. It is assumed that each frequency is proportional to both of the group totals in which it is included. For example, 43.5 per cent of the 542 children lived on farms. The number with 10 years or less schooling who lived on farms was 171. The corresponding theoretical frequency, 156, was 43.5 per cent of 358, the total children with 10 years or less of schooling.[2] Likewise, the children with over

---

[2] The theoretical frequencies may be obtained in other ways. Since 43.5 per cent of 542 children lived on farms and 66.1 per cent had 10 years or less ($358 \div 542 = 0.661$), the theoretical number for this combination was 43.5 per cent of 66.1 per cent of the total number, or 156 ($0.435 \times 0.661 \times 542 = 156$).

Likewise, this theoretical number could be obtained by multiplying the proportion with 10 years of schooling or less by the total number on farms ($0.661 \times 236 = 156$).

10 years of schooling were divided into theoretical frequencies of 80 on farms and 104 in towns. The theoretical frequencies have the same totals in both directions as the actual.

After the theoretical frequencies are obtained, χ^2 is calculated in the usual manner.

The number of frequencies in subgroups was four. However, in this case, there was only one degree of freedom, instead of three, as might be expected. The determination of any one of the four theoretical frequencies automatically fixes the other three so that the horizontal and vertical totals are the same as for the actual frequencies. In testing a two-way frequency table, the degrees of freedom are given by

$$n = (R - 1)\ (C - 1)$$

where R is the number of rows, and C the number of columns in the body of the table. In table 6, the degrees of freedom are 1 $[(2 - 1)(2 - 1) = 1]$.

The value of $\chi^2 = 7.532$ was highly significant. In other words, the distribution of former farm children now on farms according to years of schooling was different from the corresponding distribution for former farm children now in towns.[3] In short, there was a relationship between schooling and the present residence of former farm children. A higher proportion of the better-educated farm children than of those with less training move to town. That relationship was highly significant.

This relationship was tested with standard errors[4] and found to be very significant. The difference was in the method used and not in the results obtained. With standard errors, the frequencies in two-way tables of this type must be converted to *percentages* before the t test is applied. With χ^2, the relationships are tested from the *original frequencies.*

When both the t test and the χ^2 test were applied to the two-way frequency table 6, one method possessed no advantage over the other. However, it will be noted that there are only two frequencies for each classification, "years in school" and "place of residence." Two percentages may be compared as efficiently with the t test as the four frequencies are tested with chi square. However, when two-way frequency tables contain three or more classes each way, chi square has the advantage in testing all differences at once, whereas the *t* test compares only the difference between two classes.[5]

---

[3] Stated another way, the distribution of former farm children with 10 years or less of schooling according to present residence was different from the corresponding distribution for those with over 10 years of schooling. Regardless of which distributions are compared, the important point is that a relationship between schooling and present residence was indicated. That relationship was very significant.

[4] Table 12, page 343.

Use of Chi Square as a Preliminary Test

When the dependent variable is in numerical terms, simple relationships are usually shown by averages and tested by the analysis of variance. Such relationships may also be shown by two-way frequency distributions and tested by chi square. Because χ^2 is a simpler test of significance than the analysis of variance and involves much less work, a relationship may sometimes be more easily tested by χ^2 from a two-way frequency table than by the F test from averages. Consequently, chi square may have a place as a preliminary, rough test of relationships for many kinds of problems.

The relation of milk market to labor efficiency could be studied by sorting the farms according to market and averaging the index of efficiency for each type of market, with the following results:

MARKET	INDEX OF EFFICIENCY
Milk	213
Butterfat	180
None	151

In order to test the difference among these three averages with analysis of variance, the sums of squares for 707 farms must first be determined, a calculation which involves considerable busy work, especially if tabulating equipment is not available.

Another method of examining the relationship is to count the farms for several subgroups according to market and efficiency (table 7). The relationship is shown by comparison of the frequency distributions in rows or columns. The relationship is harder to see in the frequency table than in the three averages, but it is easier to test. The advantage of χ^2 is that no bothersome sums of squared deviations need be known.

The theoretical frequencies are obtained by multiplying percentages in the last line by the totals for milk markets. For example, 28.4 per cent of 353 is 100, the theoretical frequency for "milk" and "less than 140" (table 7).

Chi square, calculated by the usual method, was very significant, $\chi^2 = 78.95$.

The *frequencies* in the two-way table indicated a very significant relationship between milk market and labor efficiency. Farms selling fluid milk were more numerous as efficiency increased (54, 131, and

---

[5] For example, for testing a two-way table with three classes each way, such as table 7, the chi-square test should be used, rather than the *t* test.

TABLE 7.—TESTING RELATIONSHIPS IN A TWO-WAY FREQUENCY TABLE

RELATION OF MILK MARKET TO LABOR EFFICIENCY*

Milk market	Two-way frequency				Theoretical frequencies			
	Index labor efficiency				Index labor efficiency			
	Less than 140	140–200	More than 200	Total	Less than 140	140–200	More than 200	Total
	Number farms	*Number farms*	*Number farms*	*Number farms*	*Number farms*	*Number farms*	*Number farms*	*Number farms*
Milk	54	131	168	353	100	123	130	353
Butterfat	82	77	72	231	66	80	85	231
None	65	38	20	123	35	43	45	123
Total	201	246	260	707	201	246	260	707
Per cent	28.4	34.8	36.8	100	28.4	34.8	36.8	100

Deviations squared divided by theoretical	χ^2 for four degrees freedom† and 1 per cent probability
$\frac{(54-100)^2}{100}=21.16$	
$\frac{(82-66)^2}{66}=3.88$	
$\frac{(65-35)^2}{35}=25.71$	
$\frac{(131-123)^2}{123}=0.52$	
$\frac{(77-80)^2}{80}=0.11$	
$\frac{(38-43)^2}{43}=0.58$	
$\frac{(168-130)^2}{130}=11.11$	
$\frac{(72-85)^2}{85}=1.99$	
$\frac{(20-45)^2}{45}=13.89$	
$\chi^2=78.95$	$\chi^2=13.277$

* From table 2, page 371.

† Degrees freedom $=(R-1)\ (C-1)=(3-1)\ (3-1)=4$.

168). For farms selling no milk, the number decreased (65, 38, and 20). This indicates that farms selling milk were on the average more efficient than those selling no dairy products. This is the same relationship shown by the three averages:

MARKET	INDEX OF EFFICIENCY
Milk	213
Butterfat	180
None	151

If the relationship shown by the *frequencies* in table 7 is significant, then the same relationship shown by the three *averages* must also be significant.

The χ^2 test of *frequencies* may overstate, but usually it understates, the significance of relationship measured by the F test from *averages*. In other words, the F test is the more precise and efficient method of studying variability in averages. The χ^2 test is merely a short-cut method of approximating significance.

χ^2 WITH SMALL THEORETICAL NUMBERS

When the theoretical frequencies are smaller than ten and especially when smaller than five, the ordinary table values of χ^2, shown in table 1, page 388, are inaccurate. This is especially true when there is only one degree of freedom. It is true to a lesser extent for two or three degrees of freedom. However, the error is negligible with more than three degrees of freedom.

When there is only one degree of freedom, a simple variation in the formula for χ^2 will adjust the "calculated" χ^2 so that it is comparable with the "table" values of χ^2 in table 1, page 388. The adjustment consists of making each deviation of the actual from the estimated frequency smaller by one-half a unit.

TABLE 8.—PROBLEMS WHERE STANDARD ERRORS, ANALYSIS OF VARIANCE, AND CHI SQUARE ARE COMMONLY USED TO TEST DIFFERENCES

Test	One-way table			Two-way table			Higher-order tables
	Averages	Frequencies	Percentages	Averages	Frequencies	Percentages	
Standard error	Between two	Between two	Between two	Between two in same or different distribution	Between two in same distribution	Between two in same or different distribution	Same as two-way except more complicated
Analysis of variance	Between two, three, or all	Not used	Not used	Between two, three, or all	Not used	Not used	
Chi square	Not used	Differences among all; or differences between actual and theoretical	Not used ordinarily	Not used	Differences among all; differences between actual and theoretical or between two distributions	Not used ordinarily	

COMPARISON OF *t*, *F*, AND χ^2 TESTS OF SIGNIFICANCE IN TABULAR ANALYSIS

Standard errors, analysis of variance, and chi square may be compared on the basis of (*a*) possibilities of test, and (*b*) ease of calculation.

(*a*) In general, standard errors are applicable to frequencies, percentages, and averages; analysis of variance is applicable only to averages; and chi square, only to frequencies and indirectly to percentages (table 8).

In testing averages, either standard errors or analysis of variance can be used. However, the possibilities of analysis of variance are much greater. The analysis of variance can be used to test everything that can be tested with standard errors and, in addition, much more. Standard errors always test difference between two averages, while analysis of variance can test the whole relationship in one series of operations.

In testing frequencies and percentages, either standard errors or chi square can be used. However, if chi square is chosen, the percentages are usually converted into frequencies. When standard errors are used to test a relationship shown by a two-way frequency table, the frequencies must be converted into percentages.

The possibilities of chi square are much greater than those of standard errors. While standard error tests only the difference between two frequencies or percentages, chi square may test the differences among all frequencies in a distribution, may compare the distribution with a large variety of theoretical distributions, or may test the difference between two distributions.

Chi square and analysis of variance are much more applicable to and efficient in testing relationships than standard errors.

(*b*) From the standpoint of ease of calculation, the chi-square test is the easiest. The *t* and *F* tests are about equally difficult.

CHAPTER 22

RELIABILITY OF CORRELATION ANALYSIS

The reliability of correlation analysis will be discussed according to the following outline: (*a*) correcting measures of correlation from samples, (*b*) testing existence of correlation, (*c*) testing differences in coefficients, (*d*) testing linearity, and (*e*) testing regression coefficients.

CORRECTING MEASURES OF CORRELATION

When the number of observations is small, the gross correlation coefficient is too high and must be corrected. As the number of observations increases, the amount of revision soon becomes negligible. All formulas for calculating measures of correlation were developed under the assumption that the total population is included. In practice, however, most series of data are samples—not populations. Since the coefficient is based on a sample, it must be corrected before it can be said to apply to the population. Formulas have been devised for such corrections.

For linear partial and multiple correlations, the correction involves not only the number of observations but also the number of variables. For curvilinear correlation, the correction involves three factors—the size of the sample, the number of variables, and the number of constants in the curves.

CORRECTING GROSS CORRELATION COEFFICIENTS

For the Minneapolis price of wheat, X_1, and the United States production, X_3, the coefficient of correlation was $r_{13} = -0.469$ for the 22-year period.[1] This correlation coefficient was based on the assumption that 22 years constituted the population. Since the 22-year period was only a small sample, the correlation coefficient must be adjusted according to the following formula:

$$\bar{r}_{12}^2 = 1 - (1 - r_{12}^2)\left(\frac{N-1}{N-2}\right)$$

where $\bar{r}_{12}$ is the revised or adjusted coefficient.

---

[1] Page 187.

$$\begin{aligned}\bar{r}_{13}^2 &= 1 - [1 - (-0.469)^2]\left[\frac{22 - 1}{22 - 2}\right] \\ &= 1 - (0.780)(1.05) \\ &= 1 - 0.819 = 0.181 \\ \bar{r}_{13} &= -0.425\end{aligned}$$

The corrected, adjusted, or estimated coefficient for the population was -0.425, compared with -0.469 for the sample. The coefficient was reduced 0.044, negatively.

In the case of the correlation between the Minneapolis and Liverpool prices of wheat, $r_{12} = +0.732$, the adjusted or corrected coefficient was

$$\begin{aligned}r_{12}^2 &= 1 - [1 - (0.732)^2]\left[\frac{22 - 1}{22 - 2}\right] \\ &= 1 - (0.464)(1.05) \\ &= 1 - 0.487 = 0.513 \\ \bar{r}_{12} &= +0.716\end{aligned}$$

The corrected coefficient for the population was only a little less than that for the sample, $+0.732$. The coefficient was reduced 0.016. The adjustment of a high correlation coefficient is less than that for a low one.

If the correlation between the Minneapolis price, X_1, and the United States production, X_3, had been the same, $r_{13} = -0.469$, but based on 8 years, instead of 22 years, the corrected coefficient would have been

$$\begin{aligned}\bar{r}_{13}^2 &= 1 - [1 - (-0.469)^2]\left[\frac{8 - 1}{8 - 2}\right] \\ &= 1 - (0.780)(1.167) \\ &= 1 - 0.910 = 0.090 \\ \bar{r}_{13} &= -0.300.\end{aligned}$$

This adjusted coefficient based on 8 years' data was much less than that based on 22 years' data, $\bar{r}_{13} = -0.300$ and $\bar{r}_{13} = -0.425$, respectively. The adjustment becomes greater as the size of the sample becomes smaller.

The amount of correction or adjustment for gross correlation coefficients depends on two factors: the degree of correlation and the size of the sample.

CORRECTING MULTIPLE CORRELATION COEFFICIENTS

The formula for the adjustment of multiple correlation coefficients is as follows:

$$\bar{R}_{1\,23\cdots m}^2 = 1 - (1 - R_{1.23\cdots m}^2)\left(\frac{N - 1}{N - m}\right)$$

where $\bar{R}$ is the adjusted multiple correlation coefficient and m is the number of variables or constants in the regression equation. For the Minneapolis price of wheat and the world and United States production of wheat, X_3 and X_4, the multiple correlation coefficient[2] was $R_{1.34} = 0.658$ and the corrected coefficient, $\bar{R}_{1.34} = 0.611$ as follows:

$$\begin{aligned}
\bar{R}^2_{1.34} &= 1 - [1 - (0.658)^2]\left[\frac{22 - 1}{22 - 3}\right] \\
&= 1 - (0.567)(1.105) \\
&= 1 - 0.627 = 0.373 \\
\bar{R}_{1.34} &= 0.611
\end{aligned}$$

The adjusted coefficient for the 22 years was 0.047 less than the unadjusted.

For multiple correlation, the amount of correction depends on (*a*) the degree of correlation, (*b*) the size of the sample, and (*c*) the number of variables, m.

CORRECTING PARTIAL CORRELATION COEFFICIENTS

The adjusted partial coefficients are based on the adjusted multiple coefficients as follows:

$$\bar{r}^2_{12.34} = \frac{\bar{R}^2_{1.234} - \bar{R}^2_{1.34}}{1 - \bar{R}^2_{1.34}}$$

For the wheat problem, the unadjusted[3] and adjusted squared multiple coefficients were

$$R^2_{1.234} = 0.715 \qquad \bar{R}^2_{1.234} = 0.667$$
$$R^2_{1.34} = 0.433 \qquad \bar{R}^2_{1.34} = 0.373$$

and the adjusted partial

$$\begin{aligned}
\bar{r}^2_{12.34} &= \frac{0.667 - 0.373}{1 - 0.373} = \frac{0.294}{0.627} = 0.469 \\
\bar{r}_{12.34} &= 0.685
\end{aligned}$$

The adjusted partial coefficient was 0.021 less than the unadjusted, $r_{12.34} = 0.706$.

In partial correlation, the amount of adjustment increases as the size of the sample decreases and as the degree of correlation decreases.[4] The adjustment increases very slightly as the number of variables increases.

---

[2] Table 1, page 187.

[3] Table 1, page 187.

[4] This is most apparent in the formula for the adjusted multiple correlation coefficients.

CORRECTING INDEXES OF CORRELATION

The formula for adjusting the index of correlation is as follows:

$$\bar{\rho}^2 = 1 - (1 - \rho^2)\left(\frac{N-1}{N-m}\right)$$

where m is the number of constants in the equation. For the price and production of cabbage,[5] the unadjusted index of correlation was $\rho_{(Y=a/X^b)(LS)(\text{natural numbers})} = 0.818$. The adjusted index is

$$\begin{aligned}
\bar{\rho}^2 &= 1 - [1 - (0.818)^2]\left[\frac{20-1}{20-2}\right] \\
&= 1 - (0.331)(1.056) \\
&= 1 - 0.350 = 0.650 \\
\bar{\rho} &= 0.806
\end{aligned}$$

When rho is based on a mathematical curve, the number of constants, m, is definitely known. In this illustration, there are two constants, a and b. However, when rho is based on a freehand curve, the number of constants must be estimated. A straight line has two constants. Curves may have any number, two or more, depending on the number of bends in the curve and the smoothness of the curve between the bends. The student must estimate the constants in a curve on the basis of similar mathematical curves. In general, long, sweeping, freehand curves with one bend involve about three constants.

CORRECTING INDEXES OF MULTIPLE CORRELATION

The adjustment of indexes of correlation is the same regardless of the number of variables included.

The adjusted index for the acres of corn in North Carolina[6] is as follows:

$$\bar{\rho}^2_{1.234\ (\text{Approximation})} = 1 - (1 - \rho^2)\left(\frac{N-1}{N-m}\right)$$

where m is estimated as 7.

$$\begin{aligned}
\bar{\rho}^2_{1.234\ (\text{Approximation})} &= 1 - [1 - (0.746)^2]\left[\frac{25-1}{25-7}\right] \\
&= 1 - (0.443)(1.333) \\
&= 1 - 0.591 = 0.409 \\
\bar{\rho} &= 0.640
\end{aligned}$$

Rho was reduced by 0.106. The amount of adjustment depended on the number of observations, the number of variables, and the number of constants in the curves. The constants in the three curves[6] were estimated as follows. Since each curve was a long, sweeping, smooth curve with one bend, it was estimated to contain three constants. Individually, the three curves would have nine constants. However, when the three curves are combined into one equation, the one constant term from each curve combines into one for the whole equation leaving seven constants[7] $(9 - 1 - 1 - 1 + 1 = 7)$.

---

[5] Page 203.

[6] Figures 5, 6, and 7 on pages 225 to 227.

TESTING SIGNIFICANCE OF CORRELATION

TESTING EXISTENCE OF CORRELATION

The previous section dealt with the best estimates of correlation coefficients from samples. However, it gave no indication of the reliability of such estimates. This section deals with the significance of the correlation coefficients from samples. The problem is to test whether the correlation in the sample is due to chance fluctuations, that is, factors not considered, or to a relationship between the factors correlated.

Significance of Gross Correlation

Gross coefficients may be tested for significance in any one of three ways: (*a*) t test with standard error of r, (*b*) analysis of variance, and (*c*) t test with the standard error of z.

(*a*) *t Test with Standard Error of r.* The t test with standard error of r is based on the null hypothesis that the coefficient for the whole population is $r = 0$. Under this hypothesis, the standard error of r is

$$\sigma_r = \sqrt{\frac{1 - r^2}{N - 2}} \; .$$

The quantity $N - 2$ is the degrees of freedom[8] about the regression line $Y = a + bX$. For the price and production of wheat,[9] $r_{13} = -0.469$,

---

[7] In this problem of four variables, there are three curves which might be given by the three following equations:

$$X_1 = a_1 + b_1 f(X_2) + c_1 f'(X_2)$$
$$X_1 = a_2 + b_2 f(X_3) + c_2 f'(X_3)$$
$$X_1 = a_3 + b_3 f(X_4) + c_3 f'(X_4)$$

Each equation has three constants. In a combined equation, the constants a_1, a_2, and a_3 which determine the level of the three curves are reduced to *one* constant, a, for the one combined curve, and the six terms containing the independent variables have *six* other constants.

[8] The degrees of freedom about the arithmetic mean are $N - 1$. An arithmetic mean may be given by the equation $Y = a$, where, of course, a is the arithmetic mean, a constant. In the equation of a straight line, $Y = a + bX$, there is an additional constant, b, which is the regression coefficient which takes up another degree of freedom. If the straight line has two degrees of freedom, the remaining degrees about the straight line are $N - 2$.

[9] Page 187.

$$\sigma_{r_{13}} = \sqrt{\frac{1 - (-0.469)^2}{22 - 2}} = \sqrt{\frac{0.780}{20}} = \sqrt{0.039}$$
$$= 0.197$$

The next step is to test the null hypothesis by calculating t:

$$t = \frac{r - 0}{\sigma_r} = \frac{0.469 - 0}{0.197} = 2.38$$

The degrees of freedom, n, are $N - 2$, or 20. The 95 per cent value of t was 2.09. The correlation between the United States production and the Minneapolis price of wheat, $r_{13} = -0.469$, was significant; that is, the correlation was not entirely due to chance.[10] Since the 99 per cent value of t was 2.84, the correlation was not high enough to be termed very significant.

(*b*) *Analysis of Variance.* The sum of squared deviations in a dependent variable, $N\sigma^2$, may be divided into two parts: (*a*) that explained by the independent variable, $r^2N\sigma^2$, and (*b*) that unexplained by the independent variable, $(1 - r^2)N\sigma^2$. The squared correlation coefficient indicates the proportion of explained squared variability. Since it is the variance ratio which is desired rather than the variance, it is sufficient to deal with the proportions r^2 and $1 - r^2$ rather than the actual sums of the squares, $r^2N\sigma^2$ and $(1 - r^2)N\sigma^2$. The ratio of the two proportions is the same as the ratio of the two sums of squared deviations.

The procedure in testing a gross correlation coefficient is as follows: The proportion explained, r^2, is divided by the degrees of freedom, 1, for the one regression coefficient of a straight line. The proportion unexplained, $1 - r^2$, is divided by the degrees of freedom, $N - 2$, about the regression line. The variance ratio is calculated from the two quotients, and tested in the usual way. For testing correlation coefficients, the basis of comparison is the unexplained variability.

For the price and production of wheat,[11] $r_{13} = -0.469$ would be tested as follows:

	PROPORTION OF SQUARED DEVIATIONS	DEGREES FREEDOM	VARIANCE[12] PROPORTION	VARIANCE RATIO, F	95 PER CENT VALUE[13] OF F
Explained by X_3	$r^2_{13} = 0.220$	$1 = 1$	0.220	5.64	4.35
Unaccounted for	$1 - r^2_{13} = 0.780$	N-2 = 20	0.039←—*Basis of comparison*		
Total	$1 = 1$	N-1 = 21			

---

[10] The fact that it is significant does not prove that the association is a casual one. That can be determined only by judgment.

[11] Pages 187 and 405.

[12] Expressed as a proportion of total sum of squared deviations.

[13] Table 2, page 350.

Since the variance ratio, $F = 5.64$, was greater than the 95 per cent value of F, 4.35, the association between price and production of wheat may be said to be significant.

(*c*) *t Test with the Standard Error of z, Frequently Called the z Transformation.* For some purposes, it is necessary before calculating t to transform r into z and then test z by means of its standard error. The expression z is merely the following function of r:

$$z = 1.1513\,[\log(1 + r) - \log(1 - r)]$$

in terms of common logarithms. The standard error of z is given by

$$\sigma_z = \frac{1}{\sqrt{N - 3}}$$

To test whether a correlation coefficient is significantly greater than zero, the standard error of z is no more useful than the standard error of r. However, when r is assumed to be other than zero, estimates of its standard error are erroneous. The only value of z transformation is where the standard error of r cannot be used. With the σ_r, the correlation $r_{13} = -0.469$ could be tested to determine whether it was significantly greater than zero, but not whether it was significantly greater than -0.30, -0.20, or -0.10. With the σ_z, the latter type of test can be made as follows:

1. Calculate z for $r = -0.469$.

$$\begin{aligned} z &= 1.1513[\log(1 + r) - \log(1 - r)] \\ &= 1.1513\{\log[1 + (-0.469)] - \log[1 - (-0.469)]\} \\ &= 1.1513(\log 0.531 - \log 1.469) \\ &= 1.1513[(9.7251 - 10) - (0.1670)] \\ &= 1.1513(-0.4419) \\ &= -0.5088 \end{aligned}$$

2. Calculate z for a hypothetical r, say $r = -0.20$.

$$\begin{aligned} z &= 1.1513\{\log[1 + (-0.20)] - \log[1 - (-0.20)]\} \\ &= 1.1513(\log 0.80 - \log 1.20) \\ &= 1.1513[(9.9031 - 10) - 0.0792] \\ &= 1.1513(-0.1761) \\ &= -0.2027 \end{aligned}$$

3. Calculate the standard error of z.

$$\sigma_z = \frac{1}{\sqrt{N - 3}} = \frac{1}{\sqrt{22 - 3}} = \frac{1}{\sqrt{19}} = \frac{1}{4.359} = 0.2294$$

4. Calculate t where the hypothetical $r_{13} = -0.20$ and hypothetical $z = -0.2027$.

$$= \frac{\text{Actual } z - \text{Hypothetical } z}{\sigma_z}$$

$$= \frac{-0.5088 - (-0.2027)}{0.2294} = \frac{-0.3061}{0.2294}$$

$$= -1.33$$

This value,[14] $t = 1.33$, may be compared with 1.96, the 95 per cent value[15] of t where $n = \infty$. The correlation $r_{13} = -0.469$ is not significantly greater than -0.20. However, it was previously shown to be significantly greater than zero.

Significance of Partial Correlation

Partial coefficients may be tested for significance in any one of the three ways for testing gross coefficients.

(*a*) *The t Test with Standard Error of $r_{12.3\ldots m}$.* When the population partial coefficient is assumed to be zero, the standard error is

$$\sigma_{r_{12.3\ldots m}} = \sqrt{\frac{1 - r^2_{12.3\ldots m}}{N - m}}$$

where m is the total number of variables. For the Minneapolis and Liverpool prices, with the effect of production of wheat[16] eliminated, $r_{12.34} = +0.706$.

$$\sigma_{r_{12.34}} = \sqrt{\frac{1 - (+0.706)^2}{22 - 4}} = \sqrt{\frac{0.502}{18}}$$

$$= \sqrt{0.0279} = 0.167$$

The value of t is

$$t = \frac{0.706 - 0}{0.167} = 4.23$$

Since the 99 per cent value[17] of t for $n = 18$ is 2.88 and less than 4.23, the partial coefficient is highly significant.

(*b*) *Analysis of Variance.* The partial coefficient, $r_{12.34}$, can be calculated from two multiple coefficients[18] according to the following relationship:

$$r^2_{12.34} = \frac{R^2_{1.234} - R^2_{1.34}}{1 - R^2_{1.34}}$$

---

[14] The negative sign has no significance.

[15] Table 4, page 320. Note that in testing z the degrees of freedom are always infinite, ∞.

[16] Table 1, page 187.

[17] Table 4, page 320. The degrees of freedom for testing partial coefficients by this method are always $N - m$.

[18] Table 1, page 187.

The numerator of the expression is the difference between two multiples. This numerator expresses the proportion of the sum of squared deviations in the dependent variable which is explained by X_2 in addition to the proportions previously explained by X_3 and X_4. This additional explained variability, which is the basis of the partial coefficient, can be tested with the analysis of variance as follows:

	Proportion of Squared Deviations	Degrees Freedom	Variance[19]	Variance Ratio, F	99 Per Cent Value[20] of F
Explained by X_3 and X_4	$R^2_{1.34} = 0.4329$	2	—	—	-
Additional explained by X_2	$R^2_{1.234} - R^2_{1.34} = 0.2824$	1	0.2824	17.9	8.28
Unaccounted for	$1 - R^2_{1.234} = 0.2847$	$N - 4 = 18$	0.0158←-*Basis of comparison*		
Total	$1 = 1.0$	$N - 1 = 21$			

The size of the variance ratio F indicates that the additional variability explained by X_2 is highly significant. This is exactly the same as proving the partial coefficient $r_{12.34}$ highly significant.[21]

This test may also be used as a criterion of whether to include an additional variable in a multiple correlation analysis. If the additional effect is significant as in this example, the additional variable X_2 should probably be included with X_1, X_3, and X_4.

(*c*) *t Test with the Standard Error of z.* The expression z is calculated in exactly the same manner for partial as for gross coefficients. However, the standard error of z is slightly different.

$$\sigma_z = \frac{1}{\sqrt{N - m - 1}}$$

where m is the total number of variables. Whether the partial coefficient $r_{12.34} = +0.706$ is significantly greater than, say, $+0.40$, can be tested as follows:

1. Calculate z for $r_{12.34} = +0.706$.

$$\begin{aligned} z &= 1.1513[\log(1 + 0.706) - \log(1 - 0.706)] \\ &= 1.1513[0.2320 - (9.4683 - 10)] = 1.1513(0.7637) \\ &= 0.879 \end{aligned}$$

2. Calculate z for the hypothetical coefficient $r_{12.34} = +0.40$.

$$\begin{aligned} z &= 1.1513(\log 1.40 - \log 0.60) \\ &= 1.1513[0.1461 - (9.7782 - 10)] \\ &= 1.1513(0.3679) \\ &= 0.424 \end{aligned}$$

---

[19] Expressed as a proportion of the total sum of squared deviations.

[20] Table 2, page 350.

[21] With this method, the partial effect of X_3 was tested without calculating the partial coefficient $r_{12.34}$.

3. Calculate the standard error of z.

$$\sigma_z = \frac{1}{\sqrt{N - m - 1}} = \frac{1}{\sqrt{22 - 4 - 1}} = \frac{1}{\sqrt{17}} = \frac{1}{4.123}$$
$$= 0.2425$$

4. Calculate t for the hypothetical coefficient $r_{12.34} = 0.40$ and $z = 0.424$.

$$t = \frac{0.879 - 0.424}{0.2425} = \frac{0.455}{0.2425}$$
$$= 1.88$$

Since the 95 per cent value of t where $n = \infty$ is 1.96, the partial coefficient $r_{12.34} = +0.706$ can hardly be said to be significantly greater than $+0.40$.

Significance of Multiple Correlation

Multiple coefficients of correlation are tested with the analysis of variance in the same manner as the gross coefficients[22] under method (*b*). The proportion of sum of squared deviations explained by the independent variables is R^2 and the unexplained proportion $1 - R^2$. The degrees of freedom for R^2 are $m - 1$, or the number of independent variables. The degrees of freedom for $1 - R^2$ are $N - m$. The multiple relation of the world and United States production to the Minneapolis price of wheat,[23] $R_{1.34} = 0.658$, could be tested as follows:

	Proportion of Squared Deviations	Degrees Freedom	Variance Proportion	Variance Ratio, F	99 Per Cent Value[24] of F
Explained by X_3 and X_4	$R^2_{1.34} = 0.433$	$m - 1 = 2$	0.217	7.2	5.93
Unaccounted for	$1 - R^2_{1.34} = 0.567$	$N - m = 19$	0.030	← *Basis of comparison*	
Total	$1 = 1.0$	$N - 1 = 21$			

The size of the calculated variance ratio, $F = 7.2$, indicates that the multiple relationship was highly significant. In other words, the price of wheat was very significantly associated with the world and United States production. This does not mean that a significant relationship existed between X_1 and X_3 or between X_1 and X_4. It indicates that there was a significant relationship between X_1 and both X_3 and X_4 taken together.

---

[22] Page 406.

[23] Page 187.

[24] Table 2, page 350.

Significance of Indexes of Correlation, Gross and Multiple

Gross and multiple indexes of correlation can be tested by the analysis of variance in the same manner as linear gross and multiple coefficients of correlation.

For the production and price of cabbage, the index of gross correlation[25] was $\rho_{(Y=a/X^b)(LS)(\text{natural numbers})} = 0.818$. The explained variability, ρ^2, is compared with the unexplained variability, $1 - \rho^2$. The degrees of freedom for ρ^2 were $m - 1$, where m is the number of constants in the curve on which rho was based. The degrees of freedom for $1 - \rho^2$ were $N - m$. The steps of the test may be summarized as follows:

	Proportion of Sum of Squared Deviations	Degrees Freedom	Variance Proportion	Variance Ratio, F	99 Per Cent Value of F
Explained by X in curve $Y = \frac{a}{X^b}$	$\rho^2 = 0.669$	1	0.669	36.4	8.28
Unaccounted for	$1 - \rho^2 = 0.331$	18	0.0184 ← Basis of comparison		
Total	$1 = 1.0$	19			

The relation between the price and production of cabbage in terms of the curve $Y = \frac{a}{X^b}$ was highly significant.

The same method can be applied in testing the index of multiple correlation for the acres of corn in North Carolina,[26] $\rho_{1.234\ (\text{approximation})} = 0.746$. The degrees of freedom for ρ^2 were $m - 1$, where m was the number of constants estimated necessary to express the curves in equation form. The steps were summarized as follows:

	Proportion of Sum of Squared Deviations	Degrees Freedom	Variance Proportion	Variance Ratio, F	95 Per Cent Value of F
Explained by X_2, X_3, and X_4 freehand curves	$\rho^2 = 0.557$	$m - 1 = 6$	0.0928	3.77	2.66
Unaccounted for	$1 - \rho^2 = 0.443$	$N - m = 18$	0.0246 ← Basis of comparison		
Total	$1 = 1.0$	$N - 1 = 24$			

The curvilinear multiple relationship between the acres of corn in South Carolina and the prices of corn and cotton and stocks of corn was significant.

Summary of Testing Significance of Correlation

Three methods of testing existence of correlation have been employed: (*a*) standard error of r, (*b*) analysis of variance, and (*c*) standard error of z.

---

[25] Page 203 and page 404.

[26] Page 228 and page 404.

For gross coefficients, methods (*a*) and (*b*) are the more common. The results and amount of work involved in (*a*) and (*b*) are about the same. Method (*c*) involves more work and should be used only when the coefficient is tested against a hypothetical value other than zero.

With one exception, the above principles for gross coefficients hold for partials. The standard error of r is more applicable than analysis of variance when the partial coefficient is given and the multiples from which it is derived are not known. Analysis of variance is the more applicable when the partial is not known and multiple coefficients are given.

For multiple coefficients and indexes of correlation, the analysis of variance is the accurate test.

TABLE 1.—VALUES OF CORRELATION COEFFICIENTS FOR 95 AND 99 PER CENT PROBABILITIES AND VARIOUS DEGREES OF FREEDOM

Degrees of freedom[*] $N-m$	Gross and partial coefficients[†] r_{12} and $r_{12.3\ldots m}$		Multiple coefficients				Degrees of freedom[*] $N-m$	Gross and partial coefficients[†] r_{12} and $r_{12.3\ldots m}$		Multiple coefficients			
			Three variables $R_{1.23}$[‡]		Four variables $R_{1.234}$[§]					Three variables $R_{1.23}$[‡]		Four variables $R_{1.234}$[§]	
	Probability 95	99	Probability 95	99	Probability 95	99		Probability 95	99	Probability 95	99	Probability 95	99
1	.997	1.000	.999	1.000	.999	1.000	23	.396	.505	.479	.574	.532	.619
2	.950	.990	.975	.995	.983	.997	24	.388	.496	.470	.565	.523	.609
3	.878	.959	.930	.976	.950	.983	25	.381	.487	.462	.555	.514	.600
4	.811	.917	.881	.949	.912	.962	26	.374	.478	.454	.546	.506	.591
5	.754	.874	.836	.917	.874	.937	27	.367	.470	.446	.538	.497	.582
6	.707	.834	.795	.886	.839	.911	28	.361	.463	.439	.530	.490	.573
7	.666	.798	.758	.855	.807	.885	29	.355	.456	.432	.522	.482	.565
8	.632	.765	.726	.827	.777	.860	30	.349	.449	.426	.514	.475	.557
9	.602	.735	.697	.800	.750	.836	35	.325	.418	.397	.481	.444[‖]	.523[‖]
10	.576	.708	.671	.776	.726	.814	40	.304	.393	.373	.454	.419	.494
11	.553	.684	.648	.753	.703	.793	45	.288	.372	.353	.430	.397[‖]	.470[‖]
12	.532	.661	.627	.732	.683	.773	50	.273	.354	.336	.410	.379	.449
13	.514	.641	.608	.712	.664	.755	60	.250	.325	.308	.377	.348	.414
14	.497	.623	.590	.694	.646	.737	70	.232	.302	.286	.351	.324	.386
15	.482	.606	.574	.677	.630	.721	80	.217	.283	.269	.330	.304	.363
16	.468	.590	.559	.662	.615	.706	90	.205	.267	.254	.312	.288[‖]	.343[‖]
17	.456	.575	.545	.647	.601	.691	100	.195	.254	.241	.297	.274	.327
18	.444	.561	.532	.633	.587	.677	125	.174	.228	.216	.266	.246	.294
19	.433	.549	.520	.620	.575	.665	150	.159	.208	.198	.244	.225	.269
20	.423	.537	.509	.608	.563	.652	200	.138	.181	.172	.212	.195	.235
21	.413	.526	.498	.596	.552	.641	400	.098	.128	.122	.151	.139	.167
22	.404	.515	.488	.585	.542	.630	1000	.062	.081	.077	.096	.088	.106

[*] For gross, partial, and multiple correlations, m is the number of variables and N the number of observations. For gross correlations, m is obviously always 2.

[†] Snedecor, G. W., Statistical Methods, p. 133, 1940.

[‡] Snedecor, G. W., Statistical Methods, p. 286, 1940. [§] Computed by authors.

[‖] Based on interpolations by the authors.

Tables of Significant Gross, Partial, and Multiple Correlation Coefficients

Tables have been constructed giving the exact values of gross or partial and multiple correlation coefficients for different degrees of freedom and of probability. For example, with 20 degrees of freedom and 95 and 99 per cent probabilities, the table values of the gross coefficient r were 0.423 and 0.537 (table 1).

A correlation coefficient can be tested by comparing it with the corresponding value in such prepared tables. In the example of the production and price of wheat for 22 years,[27] there were 20 degrees of freedom and r_{13} was -0.469. Since this coefficient was more than the 95 per cent table value, 0.423, the association can be said to be significant. However, this coefficient, $r = -0.469$, was less than the 99 per cent table value, 0.537, and the association cannot be said to be highly significant.

Partial coefficients of any order can be tested in the same manner with the same set of values. For Minneapolis and Liverpool prices of wheat, with production eliminated,[28] the partial coefficient for 22 years was $r_{12.34} = 0.706$. With four variables and 22 observations, the degrees of freedom, $N - m$, were 18 $(22 - 4 = 18)$. The corresponding 95 and 99 per cent table values of the partial coefficient were 0.444 and 0.561. Since $r_{12.34} = 0.706$ was greater than the 99 per cent value, the association can be said to be very significant.

Multiple coefficients of any order can be tested in the same manner but with different sets of values for each order. The values for only two orders, $R_{1.23}$ and $R_{1.234}$, are given in table 1. For the production and Minneapolis price of wheat[29] for the 22-year period, the multiple correlation coefficient was $R_{1.34} = 0.658$. With three variables, the degrees of freedom, $N - m$, were 19 $(22 - 3 = 19)$. The corresponding 99 per cent table value of multiple correlation coefficients for three variables was 0.620. Since $R_{1.34} = 0.658$ was greater than this 99 per cent table value, the multiple correlation can be said to be very significant.

When a fourth variable, the Liverpool price,[30] X_2, was introduced, the multiple correlation coefficient was $R_{1.234} = 0.846$. For 18 degrees $(22 - 4 = 18)$ and a 99 per cent probability, the table value of the multiple correlation coefficient was 0.677. Since $R_{1.234} = 0.846$ was greater than this 99 per cent table value, the multiple relationship was highly significant.

---

[27] Pages 187 and 406.

[28] Pages 187 and 408.

[29] Pages 187 and 410.

[30] Table 1, page 187.

Since five or more variables are only rarely included in a multiple correlation problem, corresponding tables are not usually published, and the student should use the analysis-of-variance test on page 410.

Indexes of correlation can sometimes be tested with the tables of gross and multiple coefficients. When an index of gross correlation is based on a curve with two constants, the table values of the gross correlation coefficient can be used. For the 20-year curvilinear relationship between the production and price of cabbage,[31] $\rho_{(Y=a/X^b)(LS)(\text{natural numbers})} = 0.818$. Since the index was based on an equation with two constants, the table values of gross coefficients are applicable. The 99 per cent table value for 18 degrees of freedom was 0.561. Since the actual index, 0.818, was greater than the table value, this curvilinear relationship between the production and price of cabbage was highly significant.

When a curve has three or four constants, the table values of three- and four-variable multiple coefficients should be used.

For the 20-year parabolic relationship between the production and price of cabbage,[32] $\rho_{(Y=a+bX+cX^2)} = 0.862$. Since the index was based on an equation with three constants, the table values of the three variable multiples can be used. The 99 per cent table value for three variables and 17 degrees of freedom was 0.647. Since the actual index, 0.862, was greater than the 99 per cent table value, this curvilinear relation between the production and price of cabbage may also be said to be highly significant.

Indexes of multiple correlation sometimes can be compared with the table values of *multiple correlation coefficients.* When the number of constants in the curves is three or four, the table values for three and four variable multiple coefficients are applicable (table 1). However, the equations on which indexes of multiple correlation are based ordinarily involve more than four constants[33] and cannot be tested with the ordinary tables. The F test should then be used.

TESTING DIFFERENCES BETWEEN CORRELATION COEFFICIENTS

The difference between two gross or two partial correlation coefficients can be tested with the standard error of the difference between two z's.

Differences between correlation coefficients are of two distinct types:

---

[31] Pages 203 and 411.

[32] Page 205.

[33] Note that, for the index of multiple correlation for the acreage of corn in North Carolina, page 404, the number of constants was estimated to be seven. This is not an unusually large number of constants for such problems.

(*a*) Difference between two *different* coefficients in the *same* sample or problem.

(*b*) Difference between values of the *same* coefficient in two *different* samples or problems.

An example of (*a*) is the difference between r_{14}, the relation of world production to the Minneapolis price of wheat, and r_{24}, the relation of world production to the Liverpool price.[34] Both coefficients were calculated from the same sample or problem, 22 years of records. A difference between r_{14} and r_{24} would indicate whether world production was more closely related to prices at one market than to those in the other.

An example of (*b*) would be the difference between r_{14} for 1892–1913 and r_{14} for 1914–1940. In each case, r_{14} would represent the relation of world production to the Minneapolis price. A difference between the two values of r_{14} would indicate a change in the degree of relationship with the passage of time.

The two types of differences are tested by about the same methods.

Difference Between Gross Coefficients

Two *different* coefficients in the *same* problem were $r_{14} = -0.649$, for the Minneapolis price and world production of wheat,[35] and $r_{24} = -0.668$, for the Liverpool price and world production. Both coefficients apply to the same problem, 1892–1913. World production was more highly correlated with the Liverpool price than with the Minneapolis price. The question may be raised as to whether the difference, 0.019, was significant. The method for testing such differences in gross coefficients was as follows:

1. Calculate z for[36] $r_{14} = -0.649$.

$$\begin{aligned} z &= 1.1513\ (\log 0.351 - \log 1.649) \\ &= -0.774 \end{aligned}$$

2. Calculate z for $r_{24} = -0.668$.

$$\begin{aligned} z &= 1.1513\ (\log 0.332 - \log 1.668) \\ &= -0.807 \end{aligned}$$

3. Calculate the difference between the two z's.

$$D_z = -0.807 - (-0.774) = -0.033$$

---

[34] Table 3, page 192.

[35] Table 3, page 192.

[36] Page 407.

4. Calculate the standard error of the difference, D_z, between the two z's.

$$\sigma_{D_z} = \sqrt{\frac{2}{N-3}}$$
$$= \sqrt{\frac{2}{22-3}} = \sqrt{\frac{2}{19}} = \sqrt{0.1053}$$
$$= 0.324$$

5. Setting up the null hypothesis, calculate t.

$$t = \frac{D_z - 0}{\sigma_{D_z}} = \frac{0.033}{0.324} = 0.10$$

Since the 95 per cent value[37] of t for $n = \infty$ is 1.96, the difference between the two correlation coefficients is not significant.

The difference between r_{14} for 1892 to 1913 and r_{14} for 1914 to 1940, two values of the same coefficient for different samples, would be tested in the same manner as above. The only difference would be in the calculation of the standard error of the difference.

$$\sigma_{D_z} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

where the subscript 1 refers to 1892–1913; and 2, to 1914–1940.

Difference Between Partial Coefficients

Differences between two partial correlation coefficients are tested in the same manner as differences between gross. The only distinctions are in the expressions for the standard errors of the differences.

The partial coefficients, $r_{13.2} = -0.600$ and $r_{14.2} = -0.317$, measure the net effects of United States and world production, respectively, on the Minneapolis price, with the effects of the Liverpool price eliminated.[38] United States production seems to be the more closely related. Whether the difference between $r_{13.2}$ and $r_{14.2}$ is significant may be tested as follows:

1. Calculate z for $r_{13.2} = -0.600$.

$$z = 1.1513\ (\log 0.400 - \log 1.600)$$
$$= -0.693$$

2. Calculate z for $r_{14.2} = -0.317$.

$$z = 1.1513\ (\log 0.683 - \log 1.317)$$
$$= -0.328$$

---

[37] Table 4, page 320.

[38] Table 2, page 189.

3. Calculate the difference between the two z's.

$$D_z = -0.328 - (-0.693) = 0.365$$

4. Calculate the standard error of the difference between the two z's.

$$\sigma_{D_z} = \sqrt{\frac{2}{N - m - 1}}$$

$$= \sqrt{\frac{2}{22 - 3 - 1}} = \sqrt{\frac{2}{18}} = 0.333$$

5. Setting up the null hypothesis, calculate t.

$$t = \frac{D_z - 0}{\sigma_{D_z}} = \frac{0.365}{0.333} = 1.10$$

Since the 95 per cent value[39] of t for $n = \infty$ is 1.96, the difference between the two partial correlation coefficients is not significant.

The difference between the values of $r_{13.2}$ for 1892–1913 and for 1914–1940 would be tested in the same manner except that the standard error would be

$$\sigma_{D_z} = \sqrt{\frac{1}{N_1 - m_1 - 1} + \frac{1}{N_2 - m_2 - 1}}$$

where the subscript 1 refers to 1892–1913; and 2, to 1914–1940.

TESTING CURVILINEARITY

There are many problems of testing curvilinearity, some of which are as follows: (1) averages in a one-way table may be tested against the linear-regression lines; (2) a curvilinear regression may be tested against a linear regression; (3) two curvilinear regressions may be tested against each other.

Testing Relationships in One-Way Tables against Linear Regressions

Sometimes the relationship shown in a one-way table appears to be curvilinear. Such curvilinearity may be due to a truly curvilinear relationship or merely to random fluctuations in the data. Whether the relationship is significantly curvilinear can be tested with the analysis of variance as given on pages 377 to 381.

Testing Curvilinear against Linear Regressions

For the price and production of cabbage,[40] $r = -0.803$ and $\rho_{(Y=a+bX+cX^2)} = -0.862$, and their respective squares, $r^2 = 0.645$, and $\rho^2 = 0.743$

---

[39] Table 4, page 320.

[40] Page 205.

The *index* of correlation based on the curve $Y = a + bX + cX^2$ was higher than the *coefficient* of correlation based on the straight line $Y = a + bX$. This indicates that the curve fits the data better than the straight line and that the relation was curvilinear. The significance of this curvilinearity may be tested with the analysis of variance. The steps of the test may be summarized as follows:

	Proportion of Sum of Squared Deviations	Degrees Freedom	Variance Proportion	Variance Ratio, F	95 Per Cent Value of F
Explained by $Y=a+bX$	$r^2=0.645$	1	—	—	—
Additional explained by $Y=a+bX+cX^2$	$\rho^2-r^2=0.098$	1	0.098	6.5	4.45
Unaccounted for	$1-\rho^2=0.257$	17	0.0151←—*Basis of comparison*		
Total	$1=1.0$	19			

The proportion of the total variability explained by the straight line is given by r^2 with one degree of freedom. The proportion of the total variability explained by the curve is given by ρ^2 with two degrees of freedom.[41] The additional variability explained by the curve over and above that explained by the straight line is given by the difference $\rho^2 - r^2$, with one degree of freedom. This additional variability is tested against the unaccounted-for variability to determine whether the relationship is significantly curvilinear. The variance ratio was $F = 6.5$. Since the 95 per cent table value of F was 4.45, the tendency for the relationship to be curvilinear was significant.

For the acres of corn in North Carolina, the *coefficient* of multiple correlation[42] was $R_{1.234} = 0.666$, and $R^2_{1.234} = 0.444$; and the *index* of correlation by approximation[43] was $\rho_{1.234} = 0.746$, and $\rho^2_{1.234} = 0.557$. The multiple relationship was curvilinear. This was indicated by the fact that the *index* exceeded the *coefficient*. The significance of the curvilinearity can be tested as follows:

	Proportion of Sum of Squares	Degrees Freedom	Variance Proportion	Variance Ratio, F	95 Per Cent Value of F
Explained by X_2, X_3, and X_4 in linear regression	$R^2=0.444$	3	—	—	—
Additional explained by curves	$\rho^2-R^2=0.113$	3	0.0377	1.5	3.16
Unaccounted for	$1-\rho^2=0.443$	18	0.0246←—*Basis of comparison*		
Total	$1=1.0$	24			

---

[41] The degrees of freedom are always 1 less than the number of constants in the equation.

[42] Page 214.

[43] Page 228.

The additional variability explained by the three curves over and above that explained by the three straight lines is given by $\rho^2 - R^2$. Since the curves take up six degrees of freedom and the straight lines, three degrees, the difference takes three degrees. The variance ratio, $F = 1.5$, was less than the 95 per cent table value, indicating that the multiple relationship was not significantly curvilinear.

TESTING TWO CURVILINEAR REGRESSIONS AGAINST EACH OTHER

Curves with three constants often fit the relationship more closely than straight lines or curves with two constants. Likewise, curves with four or more constants often fit better than those with three. The increase in the index of correlation for a four-constant over a three-constant curve can be tested by the F test used to test the increase in ρ over r.

The difference between indexes for two curves with the same number of constants cannot be tested with this method.

Testing the difference between two curvilinear regressions is not so common a problem as testing curvilinear regressions against linear regressions or averages.

SIGNIFICANCE OF ESTIMATES BASED ON REGRESSION EQUATIONS

The greatest value of a regression equation is to estimate the dependent variable in terms of one or more independent variables. The reliability of these estimates may be tested. The standard error of an estimate based on a simple linear regression equation, $X_1 = a + b_{12}X_2$, is given by

$$\sigma_{X_1'} \quad \text{or} \quad \sigma_{1.2} = \sqrt{\sigma_1^2(1 - r_{12}^2)}$$

However, the population standard error of estimate, estimated from a sample, is

$$s_{X_1'} \quad \text{or} \quad s_{1.2} = \sqrt{\frac{N\sigma_1^2(1 - r_{12}^2)}{N - 2}}$$

For the relation of hours of labor to harvest,[44] X_1, in terms of yield of alfalfa, X_2, on 16 farms, the regression equation was $X_1 = -3.0 + 5.39X_2$, and $r_{12}^2 = 0.910$ and $\sigma_1^2 = 15.75$. The standard error of an estimate based on the equation was

$$s_{X_1'} \quad \text{or} \quad s_{1.2} = \sqrt{\frac{16(15.75)(1 - 0.910)}{16 - 2}} = \sqrt{\frac{22.68}{14}} = \sqrt{1.62}$$
$$= 1.27$$

---

[44] Pages 147 and 148.

The standard error of estimate is interpreted in about the same way as the standard error of the arithmetic mean. For 14 degrees of freedom, the 95 per cent table value of t was 2.14, and 2.14 times $s_{X'_1}$ equals 2.72 $(2.14 \times 1.27 = 2.72)$. Of all the estimates based on the regression equation, 95 per cent would be correct to within 2.72 hours. Stated another way, the chances would be 95 out of 100 that any one estimate would be correct to within 2.72 hours.

The same general method applies to estimates from linear or curvilinear regression equations or curves with any number of variables. The general formula for the population standard error of estimate is:

$$s_{X'_1} \text{ or } s_{1.23\cdots m} = \sqrt{\frac{N\sigma_1^2(1 - R^2_{1.23\cdots m})}{N - m}} \text{ or } \sqrt{\frac{N\sigma_1^2(1 - \rho^2_{1.23\cdots m})}{N - m}}$$

where m is the number of constants in the regression equation or curves.

APPENDIX A

GLOSSARY OF SYMBOLS USED IN THIS BOOK

a = a constant in an equation; average for period in trend analysis.
A = arbitrary origin.
A = arithmetic average when used with other symbols, as AX^2, AX, AX_1, AX_2, AXY, etc.; averages method.
AD = average deviation.

b = a constant in an equation; also, rate of change in regression coefficient.
b_{12} or b_{XY} = gross regression coefficient.
$b_{12.3}$ or $b_{13.24}$, etc. = net or partial regression coefficient.
$\beta_{12.3}$, $\beta_{13.24}$, etc. = small Greek letter beta = coefficient of regression in terms of the units of standard deviation; a measure of net relationships.

$$\beta_1 = \mu_3^2/\mu_2^3 = \left(\frac{\Sigma x^3}{N}\right)^2 \div \left(\frac{\Sigma x^2}{N}\right)^3.$$

$$\beta_2 = \mu_4/\mu_2^2 = \frac{\Sigma x^4}{N} \div \left(\frac{\Sigma x^2}{N}\right)^2.$$

c = a constant in an equation.
c, c_X, c_Y = correction factors for the use of arbitrary origins.

d = a constant in an equation.
d = deviations from arbitrary origin in terms of class intervals.
$|d|$ = deviations of midpoints from arbitrary origin in terms of units without regard to sign.
D = deviations of midpoints from arbitrary origin in points; difference between two statistical measures.

e = a constant in an equation.
e = 2.71828 = base of the Napierian, or natural, logarithmic system.

f = a constant in an equation.
f = a frequency; the number of observations in a given class.
f_0 = frequency of class containing mode, median, quartile, or the like.
f_o = average of two frequencies, f_1 and f_2.
f_{-1} and f_{+1} = frequencies of classes next below and next above the modal class.
f_{-i} and f_{+i} = totals of frequencies below and above class containing median, quartile, and the like.
f, f', f'' = functions of a variable.
F = variance ratio.
FH = freehand.

i = class interval.

k = coefficient of alienation.
k^2 = coefficient of non-determination.

l_{-i} and l_{+i} = lower and upper limits of class containing mode, median, or other measures.
log = logarithm.
LS = least squares.

m = midpoint.
m = number of variables.
Ma = arithmetic mean.
Ma' = hypothetical arithmetic mean.
$|Ma - A|$ = difference between arithmetic mean and arbitrary origin, without regard to sign.
Me = median.
Mg = geometric mean.
Mh = harmonic mean.
Mo = mode.
μ_4 = small Greek letter mu = $\Sigma x^4/N$.

n = degrees of freedom.
n_1, n_2 = degrees of freedom for the greater and smaller variances, respectively.
N = number of observations.
N_{-Ma} = number of items less than arithmetic mean.
N_S, N_L = number of observations whose deviations from arbitrary origin are smaller and larger than their deviations from the arithmetic mean.

O = origin on graph.

p = percentage or proportion.
p_{12} = product moment for X_1 and X_2.
p_o = average of two proportions, p_1 and p_2.
P = product sum.
P_{10}, P_{90} = tenth and ninetieth percentiles.
P.E. = probable error.

$q = (1 - p)$ where p = proportion.
$q_o = 1 - p_o$ where p_o is the average of two proportions, p_1 and p_2.
Q_1 and Q_3 = first and third quartiles.
QD = quartile deviation or semi-interquartile range.

r = rate of change plus 1.0, in compound interest equation, $Y = ar^x$.
r, r_{XY}, r_{12}, r_{23} = simple gross correlation coefficient.
r^2, r^2_{XY}, r^2_{12}, r^2_{23} = coefficient of determination.
$\bar{r}$, $\bar{r}_{XY}$, $\bar{r}_{12}$ = population correlation coefficient estimated from samples.

$r_{12.3}$, $r_{12.34}$ = partial correlation coefficient.
$\bar{r}_{12.3}$, $\bar{r}_{12.34}$ = population partial correlation coefficient estimated from samples.
${}_{12}r_{34}$ = coefficient of part correlation.
$R_{1.23}$, $R_{1.234}$ = multiple correlation coefficient.
$R^2_{1.23}$, $R^2_{1.234}$ = coefficient of determination.
$\bar{R}_{1.23}$, $\bar{R}_{1.234}$ = population coefficient of multiple correlation estimated from samples.
ρ = small Greek letter rho = index of curvilinear correlation.
$\rho_{(LS,\ Y=a/X^b)}$ = index of curvilinear correlation computed by methods of least squares about the curve $Y = a/X^b$.
$\rho_{(1.23\ \text{approximation short-cut})}$ = index of multiple correlation between three variables by short-cut approximation method.
ρ^2 = coefficient of determination.
$\bar{\rho}$ = population index of correlation estimated from a sample.

s = population standard deviation estimated from sample.
s_p = pooled estimate of population standard deviation from two or more samples.
$s_{1.2}$ = population standard error of estimate, estimated from sample.
S, S_Y, $S_{12.34}$, and $S_{\log Y}$ = standard error of estimate.
Sk = skewness.
SP = selected points.
σ = small Greek letter sigma = standard deviation.
σ^2 = variance.
σ_{Ma}, σ_f, σ_D, σ_z = standard error of mean, frequency, differences, and other statistical measures.
σ_{DMa}, σ_{Df} = standard error of the difference between two means, frequencies, and other statistical measures.
Σ = capital Greek letter sigma = summation sign.

t = number of standard errors where the number of observations is limited; ratio of a range in a statistical measure in terms of that measure's standard error.
T = number of standard errors where the number of observations is unlimited.

V, V_{AD}, V_{QD}, and V_σ = coefficient of variability based on various measures of variability.

x = deviation of the variable X from its arithmetic mean.
$|x|$ = deviation of a variable X from its arithmetic mean, without respect to sign.
X = a variable, usually independent; also, an individual variate of that variable.
X_1 = dependent variable.
X_2, X_3, and X_4 = independent variables.
X_{-Ma} = any observation less than arithmetic mean.

y = a deviation of the variable Y from its arithmetic mean.
y' = deviation of the variable Y from value estimated from a regression equation or a curve.

Y = a variable, usually dependent; also, an individual variate of that variable.
Y' = value of Y estimated from a regression equation or a curve.

z = residual, deviation of a variable from some estimated value.
z = transformation of r.
χ^2 = small Greek letter chi, squared = measure of difference between observed and theoretical frequencies.

APPENDIX B

A METHOD OF CALCULATING SUMS OF SQUARES AND SUMS OF PRODUCTS WITH TABULATING EQUIPMENT

When "Hollerith" tabulating equipment is available, sums of squares and of products may be easily obtained by a "progressive digiting" method. The procedure may be outlined as follows:

1. First, the data are punched on a tabulating card.[1] In the problem of the production and price of wheat, the information punched on the card included the year and four variables, prices of wheat at Minneapolis and Liverpool and production of wheat in the United States and in the world (table 1, left). Whenever possible, the data in their original form should be punched on the cards. However, in this example, the information consisted of first differences with

TABLE 1.—ORIGINAL DATA CODED FOR PUNCHING ON A TABULATION CARD

CHANGES* IN PRICES OF WHEAT AT MINNEAPOLIS, X_1, AND AT LIVERPOOL, X_2; IN PRODUCTION OF WHEAT IN THE UNITED STATES, X_3, AND IN THE WORLD, X_4

Original data					Coded data				
Year	X_1	X_2	X_3	X_4	Year	X_1 (+50)	X_2 (+50)	X_3 (+50)	X_4 (+50)
1892	−18	−29	− 7	+15	1892	32	21	43	065
1893	− 7	−12	−10	+ 8	1893	43	38	40	058
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
1913	+ 2	− 6	+ 2	+22	1913	52	44	52	072
Column on tabulating card.............					1, 2, 3, 4	5,6	7,8	9,10	11,12,13

* Table 1, page 170.

---

[1] In many problems, this step has already been performed for another purpose—tabular analysis. Tabulating equipment is probably more useful for tabular than for correlation analysis. Where tabular analysis is used as well as correlation analysis, the data would already be on cards when the sums of squares and/or products were desired.

both plus and minus signs. Since the tabulating machine manipulates a series of numbers all with the same sign more easily than one with both plus and minus signs, the minus signs were removed from each of the four series of first differences. This was accomplished by adding 50 to each first difference (table 1, right).

On the tabulating card, the year was placed in columns 1 to 4. The Minneapolis price of wheat, X_1, a series of two-digit numbers, occupied columns 5 and 6; X_2 occupied columns 7 and 8; X_3, columns 9 and 10; and X_4 was placed in three columns, 11, 12, and 13, because there was one year with three digits.

The coded data in table 1, right, were punched on tabulating cards with the usual equipment. The data for each year were placed on one card, and there were 22 cards in all.

2. The next step consists of grouping the years according to one variable and obtaining cumulative totals for all four variables.

The 22 cards were first sorted on column 6, the "units" column of X_1. The resulting 10 groups[2] of cards were then arranged in order so that the pack began with those punched "9" in column 6 and ended with those punched "0." Then, with the tabulating machine, cumulative totals of the four variables, X_1, X_2, X_3, and X_4 were obtained for all the different digits of column 6.

The total of X_1 for all the "9's" in column 6 was 157 (table 2, upper). The corresponding totals for X_2, X_3, and X_4 were 166, 156, and 148, respectively. The cumulative total of X_1 for all the "9's" and "8's" in column 6 was 185. Since there were no "7's" in column 6, the cumulative total for "7" was also 185. The final cumulative total of X_1 for the "0's" was 1,104.

The 22 cards were next sorted on column 5, the "10's" column of X_1. The resulting groups were then arranged in order from largest to smallest. Then, with the tabulating machine, cumulative totals of X_1, X_2, X_3, and X_4 were obtained for all the different digits of column 5.

The total of X_1 for all the "7's" in column 5 was 217 (table 2, lower). The final cumulative total[3] of X_1 was again 1,104.

3. The next step consisted of adding the columns of cumulative totals to

---

[2] Actually, there were only nine groups because there were no "7's." However, when there are no cards for one digit, a blank card is inserted in that digit's position. In this case, a blank card was placed between the "8's" and the "6's." This rule does not apply to the zero group nor does it apply to higher groups than the highest one present. In the lower part of table 2, there were no "9's," "8's," "1's," or "0's." Blank cards were not inserted for the "9's,", "8's," or "0's," but a card was inserted for the missing "1's."

[3] Regardless of which column of which variable is the basis of the sort, the last cumulative total of any variable is always the same because it is merely the sum of all values of that variable. The total of X_1 for the 22 years was 1,104 regardless of whether the cards were sorted on columns 5 or 6 (table 2, upper and lower). Likewise, the total of X_4 was always 1,283 regardless of whether the cards were sorted on X_1, columns 6 or 5; X_2, columns 8 or 7; X_3, columns 10 or 9; or X_4, columns 13, 12, or 11 (tables 2 and 3). This reappearance of the same final totals from one tabulation to the next gives a good check on the accuracy of the machine and operator.

TABLE 2.—CALCULATION OF THE SUM OF SQUARES AND OF THE SUMS OF PRODUCTS INVOLVING X_1

From Cumulative Totals* Obtained after Sorting on the Units and Tens Digits of X_1 and Tabulating X_1, X_2, X_3, and X_4

Basis of sort X_1	Cumulative totals			
	X_1	X_2	X_3	X_4
Column 6	*Columns 5–6*	*Columns 7–8*	*Columns 9–10*	*Columns 11–13*
9	157	166	156	148
8	185	214	217	227
†	185	214	217	227
6	231	272	267	266
5	411	436	471	523
4	593	599	609	721
3	805	832	812	895
2	983	989	1014	1181
1	1054	1046	1054	1215
0	~~1104~~ = ΣX_1	~~1092~~ = ΣX_2	~~1107~~ = ΣX_3	~~1283~~ = ΣX_4
Column 5				
7	217*0*	201*0*	138*0*	81*0*
6	345*0*	313*0*	233*0*	209*0*
5	722*0*	669*0*	578*0*	605*0*
4	996*0*	954*0*	872*0*	942*0*
3	1028*0*	975*0*	915*0*	1007*0*
2	1104*0*	1092*0*	1107*0*	1283*0*
‡	1104*0*	1092*0*	1107*0*	1283*0*
Adding-machine totals	*59764* = ΣX_1^2	*57728* = $\Sigma X_1 X_2$	*54317* = $\Sigma X_1 X_3$	*59503* = $\Sigma X_1 X_4$

* All figures except those in italics were recorded by the tabulating machine. The zero to the right of the numbers in the lower part of the table and the adding-machine totals were inserted.

† There were no "7's." A blank card was inserted between the "8's" and "6's."

‡ There were no "1's." A blank card was inserted after the "2's."

obtain sums of squares and sums of products. First, all cumulative totals for "0" groups were crossed out, for they should not be included. Also, all the cumulative totals for the sort on the "10's" column (column 5) were multiplied by 10. This was done by adding a zero to the right of all the values in the lower part of table 2. Finally, the cumulative totals for X_1 consisted of 16 numbers, the first being 157 and the last 11,040. The 16 numbers were summed with an adding machine and the total, 59,764, was set under the column. The totals of the X_2, X_3, and X_4 columns were obtained in the same manner.

The sums of cumulative totals were sums of squares and of products. Each total was the sum of the products of X_1 times the particular variable tabulated. The sum at the bottom of the X_1 column, 59,764, was ΣX_1X_1 or ΣX_1^2. The other sums in order were ΣX_1X_2, ΣX_1X_3, and ΣX_1X_4 (table 2, lower).

The sums of the four variables themselves had been automatically determined in the above procedure. The sum of each variable is simply the last cumulative total for that variable in any sort. For example, ΣX_1 is 1,104, the last cumulative total of X_1 from the sort on column 6.

At this point the values of ΣX_1, ΣX_2, ΣX_3, ΣX_4, ΣX_1^2, ΣX_1X_2, ΣX_1X_3, and ΣX_1X_4 had been determined (table 2). To obtain other sums of squares, ΣX_2^2, ΣX_3^2, and ΣX_4^2, and other sums of products, ΣX_2X_3, ΣX_2X_4, and ΣX_3X_4, it was necessary to repeat the above process, sorting on X_2, X_3, and X_4 in turn, and tabulating those variables.

In table 3, left, the basis of the sort was X_2, and the variables tabulated were X_2, X_3, and X_4. The sums of cumulative totals under the X_2, X_3, and X_4 columns were ΣX_2^2, ΣX_2X_3, and ΣX_2X_4, respectively.

In table 3, center, the basis of sort was X_3, and the variables tabulated were X_3 and X_4. The sums of cumulative totals were ΣX_3^2 and ΣX_3X_4.

In table 3, right, the basis of sort was X_4, and the one variable tabulated[4] was also X_4. The sum of the cumulative totals was ΣX_4^2.

All the sums, sums of squares, and sums of products may be summarized from tables 2 and 3 as follows:

X_1 SORT		X_2 SORT	X_3 SORT	X_4 SORT
$\Sigma X_1 = 1{,}104$	$\Sigma X_1^2 = 59{,}764$	$\Sigma X_2^2 = 57{,}878$	$\Sigma X_3^2 = 57{,}293$	$\Sigma X_4^2 = 87{,}767$
$\Sigma X_2 = 1{,}092$	$\Sigma X_1X_2 = 57{,}728$	$\Sigma X_2X_3 = 54{,}746$	$\Sigma X_3X_4 = 67{,}235$	
$\Sigma X_3 = 1{,}107$	$\Sigma X_1X_3 = 54{,}317$	$\Sigma X_2X_4 = 59{,}076$		
$\Sigma X_4 = 1{,}283$	$\Sigma X_1X_4 = 59{,}503$			

Any sum of products could have been determined in another way. The value of ΣX_1X_3 was obtained by sorting on X_1 and tabulating X_3. However, it could have been obtained by sorting on X_3 and tabulating X_1. The answer would have been the same. The sums of the four variables themselves, which were taken from table 2, could have been taken from any other tabulation of that particular variable. For example, since X_4 appeared in every tabulation, the sum, $\Sigma X_4 = 1{,}283$, appears nine times in tables 2 and 3.

When the data are not coded, the sums of squares and of products by this method are identical with those by other methods. When data are coded, as in the present example, the results are not the same. However, the sums of squares or of products are not the final objects of the calculations. The product

---

[4] While X_1, X_2, and X_3 were all two-digit numbers requiring two columns on the tabulating card, one value of X_4 was over 99, requiring three columns on the card for X_4. Note that there are three sets of cumulative totals based on the X_4 sort, those for the sort on the "units" column 13, the "10's" column 12, and the "100's" column 11. The totals for the column 11 sort were multiplied by 100 before adding the 19 cumulative totals together to obtain ΣX_4^2.

TABLE 3.—CALCULATION OF SUMS OF SQUARES AND SUMS OF PRODUCTS INVOLVING X_2, X_3, AND X_4

Sort X_2	Cumulative totals			Sort X_3	Cumulative totals		Sort X_4	Cumulative totals
	X_2	X_3	X_4		X_3	X_4		X_4
Column 8	*Columns 7–8*	*Columns 9–10*	*Columns 11–13*	*Column 10*	*Columns 9–10*	*Columns 11–13*	*Column 13*	*Columns 11–13*
9	49	44	35	9	186	125	9	187
8	289	320	366	8	234	181	8	517
7	450	462	503	7	281	253	7	544
6	562	554	591	6	413	453	6	696
5	607	608	687	5	523	566	5	796
4	651	660	759	4	621	697	4	874
	651	660	759	3	813	953	3	927
2	817	816	907	2	865	1025	2	1143
1	1052	1062	1239	1	977	1152		1143
0	~~1092~~	~~1107~~	~~1283~~	0	~~1107~~	~~1283~~	0	~~1283~~
Column 7				*Column 9*			*Column 12*	
7	780	590	270	6	2580	3480	9	960
6	2670	1970	1510	5	6300	7510		960
5	6960	5960	5850	4	10290	12350	7	4610
4	10150	9580	10320	3	11070	12830	6	6630
3	10530	9980	10900		11070	12830	5	8800
2	10740	10410	11550		11070	12830	4	9720
1	10920	11070	12830				3	10800
							2	12830
								12830
							Column 11	
							1	12800
							0	~~1283~~
Adding-machine totals	*57878* $= \Sigma X_2^2$	*54746* $= \Sigma X_2 X_3$	*59076* $= \Sigma X_2 X_4$	Adding-machine totals	*57293* $= \Sigma X_3^2$	*67235* $= \Sigma X_3 X_4$	Adding-machine totals	*87767* $= \Sigma X_4^2$

moments and standard deviations calculated from these sums are the results desired. Product moments, p_{12}, p_{13}, and the like, and standard deviations, σ_1^2 and the like, are exactly the same regardless of whether the data are coded, provided that the coding is done by addition or subtraction. For example, from the above values,

$$AX_2X_3 = \frac{54{,}746}{22} = 2{,}488.45 \quad AX_2 = \frac{1{,}092}{22} = 49.6364 \quad AX_3 = \frac{1{,}107}{22} = 50.3182$$

According to the usual formula,[5]

$$\begin{aligned} p_{23} &= AX_2X_3 - (AX_2)\,(AX_3) \\ &= 2{,}488.45 - (49.6364)\,(50.3182) \\ &= -9.16 \end{aligned}$$

This is exactly the same as the value of p_{23} obtained from uncoded data by another method.[6]

When tabulating machines are available, the progressive-digiting method of obtaining sums of squares and of products often saves much time and labor over other methods. The amount of time saved depends on (1) the number of observations, (2) the number of variables to be squared and cross-multiplied in the same problem, and (3) whether the data have already been punched on cards.

When the number of observations is greater than 50 or 60, the tabulating-machine method usually means a considerable saving of time. When the number of observations is small, say 15 to 20, other methods probably take less time than the tabulating-machine method.

The advantage of the tabulating-machine method increases with the number of variables. If a person were studying only the effect of X_2 on X_1, he would probably not save much time with tabulating equipment even if he had a very large number of observations. However, if he were studying the relationships among 10 factors, he would probably save time by using the machine method with only 20 to 30 observations.[7]

When data have already been punched on cards for another reason, the machine method saves even more time over other methods.

Another advantage of the tabulating-machine method in addition to the saving of time is greater accuracy. This advantage is small when there are only 15 to 20 observations, but becomes greater as the number of observations increases.

The chief limitations of the machine method concern the unfamiliarity of the statistician with the machines and the machine operator with the method.

[5] Page 172.

[6] Page 172.

[7] In this problem of price and production of wheat, it is doubtful whether the tabulating-machine method took any less time than the usual methods. There were only 4 variables and only 22 observations (years). This example was used primarily because it had appeared before in the chapters on multiple and partial correlation, pages 169 and 186.

APPENDIX C

THE DOOLITTLE METHOD OF SOLVING NORMAL EQUATIONS FOR NET REGRESSION COEFFICIENTS

The method used for solving normal equations[1] in chapter 10 is a general method applicable to various types of simultaneous equations. However, the unique nature of the normal equations makes possible another procedure known as the Doolittle method. For the beginner, the method given in chapter 10 is probably the easier to follow. For the worker with many sets of equations to solve, the Doolittle method probably saves time.

The normal equations used to illustrate the Doolittle method were the same as those on page 173.

$$\begin{array}{lrrrr}
\text{(I)} & +167.0496b_{12.34} & -\ \ 9.1570b_{13.24} & -209.4300b_{14.23} & = +133.1570 \\
\text{(II)} & -\ \ 9.1570b_{12.34} & +\ \ 72.3078b_{13.24} & +121.6713b_{14.23} & = -\ \ 56.1033 \\
\text{(III)} & -209.4300b_{12.34} & +121.6713b_{13.24} & +588.3984b_{14.23} & = -221.8304
\end{array}$$

However, with the Doolittle method, the first term in equation II, $-9.1570b_{12.34}$, and the first and second terms in equation III are not used. Another difference is that the three constant terms on the right side of the equation are transferred to the left of the "equals" symbols; and, of course, their signs are changed.

The Doolittle method proceeds as on page 432.

---

[1] Pages 173 to 175.

DOOLITTLE METHOD OF SOLVING NORMAL EQUATIONS

Line	Procedure	Calculations
1.	Equation I	$+167.0496b_{12.34}-9.1570b_{13.24}-209.4300b_{14.23}-133.1570=0$
2.	Equation II	$+72.3078b_{13.24}+121.6713b_{14.23}+56.1033=0$
3.	Equation III	$+588.3984b_{14.23}+221.8304=0$
4.	Write equation I	$+167.0496b_{12.34}-9.1570b_{13.24}-209.4300b_{14.23}-133.1570=0$
5.	Line 4 ÷ +167.0496, signs changed	$-1.0b_{12.34}+0.054816b_{13.24}+1.253700b_{14.23}+0.797111=0$
6.	Write equation II	$+72.3078b_{13.24}+121.6713b_{14.23}+56.1033=0$
7.	Line 4 × +0.054816, $b_{12.34}$ term omitted	$-0.5020b_{13.24}-11.4801b_{14.23}-7.2991=0$
8.	Line 6 + line 7	$+71.8058b_{13.24}+110.1912b_{14.23}+48.8042=0$
9.	Line 8 ÷ +71.8058, signs changed	$-1.0b_{13.24}-1.534572b_{14.23}-0.679669=0$
10.	Write equation III	$+588.3984b_{14.23}+221.8304=0$
11.	Line 4 × +1.253700, $b_{12.34}$ and $b_{13.24}$ terms omitted	$-262.5624b_{14.23}-166.9389=0$
12.	Line 8 × −1.534572, $b_{12.34}$ and $b_{13.24}$ terms omitted	$-169.0963b_{14.23}-74.8936=0$
13.	Line 10 + line 11 + line 12	$+156.7397b_{14.23}-20.0021=0$
14.	Line 13 ÷ +156.7397, signs changed	$-1.0\,b_{14.23}+0.127613=0$
15.	Value of $b_{14.23}$	$+1.0\,b_{14.23}=+0.127613$
16.	Line 9 with value of $b_{14.23}$ substituted	$-1.0b_{13.24}-1.534572(+0.127613)-0.679669=0$
17.	Simplification	$-1.0b_{13.24}-0.875500=0$
18.	Value of $b_{13.24}$	$+1.0b_{13.24}=-0.875500$
19.	Line 5 with values of $b_{13.24}$ and $b_{14.23}$ substituted	$-1.0b_{12.34}+0.054816(-0.87550)+1.253700(+0.127613)+0.797111=0$
20.	Simplification	$-1.0b_{12.34}+0.909108=0$
21.	Value* of $b_{12.34}$	$+1.0b_{12.34}=+0.909108$

Check:

Substitute in equation I the values of $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$.

$$+167.0496(+0.909108)-9.1570(-0.875500)-209.4300(+0.127613)-133.1570=0$$
$$+151.8661+8.0170-26.7260-133.1570=0$$
$$+0.0001=0$$

---

* Lines 15, 18, and 21 are not absolutely necessary, because the values of $b_{14.23}$, $b_{13.24}$, and $b_{12.34}$ are also given to the left of the equals signs in lines 14, 17, and 20, respectively.

APPENDIX D

GENERAL INFORMATION FOR TABLES IN CHAPTER 15

RELATION OF SIZE OF FARM, CROP YIELDS, AND LABOR EFFICIENCY TO INCOME ON 907 NEW YORK FARMS, 1927

Size of farm, X_2			Crop index, X_3			Labor efficiency, X_4			Income, X_1		Number farms
Group	Total	Average	Group	Total	Average	Group	Total	Average	Total	Average	
Small	28,530	157.6	Poor	13,235	73.1	Low	22,750	125.7	\$−21,500	\$−119	181
"	11,390	232.4	"	3,711	75.7	High	10,860	221.6	+18,800	+384	49
"	28,070	169.1	Good	19,623	118.2	Low	21,810	131.4	+16,700	+101	166
"	10,620	230.9	"	5,334	116.0	High	10,770	234.1	+16,600	+361	46
Large	15,830	351.8	Poor	3,716	82.6	Low	7,110	158.0	−12,200	−271	45
"	75,790	438.1	"	13,319	77.0	High	51,560	298.0	+102,500	+592	173
"	24,840	365.3	Good	7,987	117.5	Low	10,820	159.1	+15,800	+232	68
"	90,090	503.3	"	20,658	115.4	High	51,270	286.4	+203,800	+1,139	179
Total	285,160	2,448.5		87,583	775.5		186,950	1,614.3	+340,500	+2,419	907
Averages											
Simple	—	306.1		—	96.9		—	201.8	—	+302.4	—
Weighted	314.4	—		96.6	—		206.1	—	+375.4	—	—

INDEX

A

Additive relationships, 128, 246, 249
  analysis of variance applied, 374
  and joint compared, 261
  tested in three-way table, 383, 386
Aggregative, 56, 66
Alienation, coefficient of, 160
Analysis of variance, 345
  applications, to correlation, curvilinear, 411
      gross, 406
      multiple, 410
      partial, 408
      testing curvilinearity, 417
    to tabulation, additive relationships, 374
      curvilinear relationships, 377
      difference between two means, 354
        t test *vs.* *F* test, 356
      joint relationships, 374
      non-numerical variables, 360
      one-way classification, 357
        consistency of relationships, 358
      three-way tables, 381
      two-way classifications with equal subgroups, 360
        more than one observation in each subgroup, 360
        one observation in subgroup, 365
      two-way tables with unequal subgroups, 370
  compared with standard errors and chi square, 399
  experimental error, 355
  F and *t* test compared, 368
Arbitrary origin, 18, 47, 155
Arithmetic mean, *see* Mean
Asymmetrical distribution, 11
Average deviation, *see* Deviation, average
Averages, *see* Mean
  various averages compared, 34
Averages method for linear trend, 78

B

Base periods, 73
Bauman, A. O., 98
Bean, L. H., 230, 244
Benner, C. L., 182
Bennett, K. R., 183, 211, 330
Beta coefficient, 199
Bias, 301, 302
Bivariate frequency distribution, 155
Black, J. D., 162, 244
Bowley, A. L., 140
Brandow, G. E., 88

C

Campbell, C. E., 182, 244
Carmichael, F. L., 98
Cassetta, Mrs. J. V., v
Central tendency, *see* Mean; Median; Mode
Chi square (χ^2), 387
  application, 389
    as preliminary test, 397
    differences in one-way frequency table, 393
    sample compared with population, 391
    testing relationships in one-way frequency table, 393
    two-way frequency table, 395
  compared with standard errors and analysis of variance, 399
  table of values, 388
  with small numbers, 398
Classes, interval, 4
  location of limits, 5
  number, 2
  size, 4
  unequal, 4
Coefficient, of correlation, *see* Correlation
  of regression, *see* Regression

- Correction factor, averages, 19
  - correlation, 156
  - standard deviation, 48
- Correlation, additive, and joint compared, 261
    - and joint relationships, 291
      - compared with tabulation, 293
  - advantages and disadvantages, 266, 273, 276, 282, 293, 295–299
  - causation, 194–195
  - corrected values, 401–405
    - gross, 401
    - indexes, of correlation, 404
      - of multiple correlation, 404
    - multiple, 402
    - partial, 403
  - curvilinear, multiple, *see* Correlation, multiple, index of
    - simple or gross, 200, 269
      - advantages and disadvantages, 210
      - effect of extreme residuals, 205
      - effect of flexibility of curves, 206
      - effect of method of fitting curves, 206
      - effect of method of measuring residuals, 207
      - from different curves, 203
      - index of correlation, 200
      - rho (ρ), 200
        - calculation of, 202
        - characteristics of, 210
        - from different curves, 203
        - importance of defining, 208
        - problems in choosing, 209
      - uses, 211
  - gross, linear, 144–150
      - advantages and disadvantages, 161
      - coefficient, 150
        - compared with determination and non-determination, 163
      - compared with tabulation, 266
      - double-entry table, 155
      - meaning of, 143
      - methods, advantages and disadvantages, 161
        - compared, 159
        - least-squares, 146
- Correlation—(*Continued*)
  - gross, linear—(*Continued*)
      - methods—(*Continued*)
        - product-moment, with deviations, 150
          - with grouped data, 155
          - without deviations, 153
      - product moment, 151, 169, 172
      - product sum, 153
      - regression equation, 160
        - tabular use of, 164
      - simple explanation, 144
  - interserial, 194, 276
    - compared with tabulation, 276
  - joint, 246, 291
    - advantages and disadvantages, 260
    - and additive compared, 261
    - approximation method, 252
      - calculation of rho, 255
    - compared with tabulation, 293
    - comparison of additive and joint, 249
    - curvilinear, 250
    - least-squares method, 246
      - and approximation compared, 260
      - limitations of, 251
    - linear, 246
    - presentation of joint relationships, 257
      - graphic, 258
      - graphic and tabular compared, 257
      - tabular, 257
    - regression equations, 247
    - rho, 250
    - tabular presentation, 251
    - two-dimensional graph, 252
      - contour lines, 254
    - uses, 262
    - with more than two independent variables, 260
  - linear gross, *see* Correlation, gross, linear
  - multiple, curvilinear, 212
    - index of, 212
      - advantages and disadvantages, 243
      - approximation analysis, 217
        - from linear multiple regression, 217
          - advantages and disadvantages, 243

Correlation—(*Continued*)
  multiple—(*Continued*)
    index of—(*Continued*)
      approximation analysis—(*Continued*)
        short-cut method, 230
          advantages and disadvantages, 244
          guide to drawing approximations, 239
      characteristics of curvilinear methods, 241
      comparison of curvilinear methods, 242
      least-squares analysis, 213
        advantages and disadvantages, 243
      rho, 215, 228, 236
    linear, advantages and disadvantages, 178–180
      determination of, 168
      from partial, 195
      graphic presentation of results, 178
      meaning, 166
      partial regression coefficient, 163
      product moment, 169, 172
      products and squares for, 169
      R, calculation of, 175
        interpretation of, 176
      regression equation, 176
        coefficients, 177
        tabular use of, 183
      simultaneous equations for, 173
      standard error of estimate, 167
      uses, 181
    procedures, 263
    relationships, 3 variables, 273
        compared with tabulation, 273
      4 variables, 282
        compared with tabulation, 282
  part, 198
  partial, 185
    characteristics, 196
    compared with gross, 190
    first-order coefficients, 189, 191
    from gross correlations, 191
    from multiple correlations, 185
    interpretation, 196
    interserial correlation, 194
    limitations, 196

Correlation—(*Continued*)
  partial—(*Continued*)
    second-order coefficients, 185, 194
    uses, 197
  simple, *see* Correlation, curvilinear; Correlation, gross
  testing significance, curvilinear correlation, 411
    curvilinearity, 417
    difference between two coefficients, 414
      gross, 415
      partial, 416
    estimates based on regression, 419
    gross correlation, 405, 413
    multiple correlation, 410, 413
    partial correlation, 408, 413
    significant values, 412
Cox, R. W., 103, 181, 182
Cumulative chart, 9
Curve fitting, *see* Least-squares method; Secular trend
Curvilinear correlation, *see* Correlation, curvilinear
Curvilinear relationships, analysis of variance applied, 377
Curvilinearity, analysis of variance applied, 417
Cycles, annual, methods for, comparison, 115
      first-differences, 105
      percentage-of-moving-averages, 108
      percentage-of-preceding-year, 106
      percentage-of-straight-line-trend, 107
      purchasing-power, 109
    uses, 111
  monthly, methods for, moving-average, 117
      percentage-of-corresponding-month, 116
      purchasing-power, 117
    uses, 119

D

Davenport, E., 162
Deciles, 40
Deflated series compared with purchasing power, 109

DeGraff, H. F., 128
Degrees of freedom, 321
  for chi square, 389
    subdivision of variability, 347
    variance, 347
    various *t* tests, 322
Dependent variable, *see* Variables
Determination, coefficient of, 149, 176
Deviation, average, 40
    advantages and disadvantages, 49
    coefficient of variability, 43
    compared with standard and quartile, 49
    from grouped data, 42
    from ungrouped data, 40
    standard error of, 316
    standard error of difference between, 316
    uses, 51
  mean, *see* Deviation, average
  quartile, 38
    advantages and disadvantages, 49
    compared with average and standard, 49
  standard, 44
    advantages and disadvantages, 49
    coefficient of variability, 49
    compared with average and quartile, 49
    from grouped data, 46
    from ungrouped data, 44
    in means, 304
      from population, 308
      from sample, 308
    pooled, 312
    standard error of, 316
    standard error of difference between, 316
    uses, 53
Differences, first, 105, 284
  paired, 336
    standard error of mean for, 337
  second, 284, 285
    standard error of, 332
  third, 286
  weighted, 130
Discrepance, 362, 366, 372
Dispersion, 36
Distributions, *see* Frequency distributions
Doolittle method, 431
Double-classification table, 155
Double-entry tables, 155

E

Elliott, F. F., 244
Enström, A. F., 88
Error, *see* Standard error
Experimental error, 355
Exponential curve, 83
Ezekiel, M., 140, 162, 185, 217, 242, 244

F

F, variance ratio, 348, 349
    F test *vs. t* test, 368
    table of, 350–353
Falkner, H. D., 98
Findlen, P. J., 136
Fish, M., 211
Fisher, R. A., v, 320
Frequencies, standard error of, 314
  standard error of differences between, 314
Frequency distributions, 1
  asymmetrical, 11
  bivariate, 155
  class interval, 4
  comparison of, 13
  cumulative, 8
  graphic representation, 9
  J-shaped, 11
  location of class limits, 5
  multi-modal, 11
  number of classes, 2
  ogive, 9, 10
  relative, 8
  size of classes, 4
  skewed, 11
  symmetrical, 10
  unequal classes, 4
  uses, 15
  U-shaped, 13
Frequency table, one-way, 2
    t test, 339
    χ^2 test, 393
  two-way, 137, 156
    t test, 341
    χ^2 test, 395

G

Gabriel, H. S., 182
Galton, F., 147
Gans, A. R., 182
Geometric mean, *see* Mean
Goodness of fit, chi-square test for, 389
Goulden, C. H., v, 320
Gross correlation, *see* Correlation

H

Haas, G. C., 182
Harmonic, *see* Mean
Harper, F. A., 130
Heflebower, R. B., 182
Himmel, J. P., 340
Histogram, 9
Hitchcock, J. A., 181, 182

I

Index numbers, base periods, 73
  comparison, 67
  defined, 55
  effect of method, weighting, and type of commodity, 74
  time-reversal test, 70
  types of commodities, 73
  unweighted, 56
    arithmetic mean of relatives, 58
    geometric mean of relatives, 60
    median of relatives, 59
    sum of numbers or simple aggregative, 56
  weighted, 62
    aggregative, 66
    arithmetic mean of relatives, 62
      multipliers, 64
    geometric mean, 65
  weights, determination of, 62, 71
    variable, 72
Index of multiple correlation, *see* Correlation, multiple, index of
Inference, statistical, 301, 307
Interpolation of median, 24
Interserial correlation, 194, 276

J

Jesness, O. B., 126
Joint correlation, *see* Correlation
Joint relationships, 132, 246, 249, 283
  analysis of variance applied, 374
  and additive compared, 261
  correlation and tabulation compared, 293
  significance of, 335
  tested in three-way table, 383, 386
Jones, D. C., 140
Jordan, E. M., 135

K

Kincer, J. B., 181, 182
Koller, E. F., 126
Kurtosis, 36, 54

L

LaMont, T. E., 137
Least-squares method, for curvilinear correlation, 206, 213
  for cycles, 107
  for gross correlation, 146
  for joint correlation, 246
  for secular trend, 79
Leptokurtic, 54
Linear correlation, *see* Correlation, gross; Correlation, multiple; Correlation, partial
Linear trends, 76
Link-relative method, 95

M

Macaulay, F. R., 98
McCormick, T. C., 342
Malenbaum, W., 244
Mattice, W. A., 181, 182
Mean, arithmetic, advantages and disadvantages, 22
    analysis of variance for difference between, 354
    arbitrary origin, 18, 47, 155
    characteristics, 22
    effect of shifting arbitrary origin, 22

Mean, arithmetic—(*Continued*)
    effect of size of class interval, 22
    from frequency distribution, 16
    from individual items, 16
    hypothetical, 317
    standard error of, 304–307
    standard error of difference between two means, 311
    standard error of second difference, 313
    uses, 23
  geometric, 30
    advantages and disadvantages, 31
    characteristics, 31
    uses, 32
  harmonic, 32
    advantages and disadvantages, 33
    characteristics, 33
    used in analysis of variance, 373
    uses, 33
Mean deviation, *see* Deviation, average
Median, 23
  advantages and disadvantages, 26
  characteristics, 26
  determination of, from frequency distributions, 24
    from individual items, 23
  from cumulative polygon, 25
  standard error of, 315
  standard error of difference between, 315
  uses, 26
Mesokurtic, 54
Mills, F. C., 54, 301
Miner, J. R., 191
Mode, 27
  advantages and disadvantages, 29
  approximation, 28
  characteristics, 29
  uses, 29
Moving averages, 93, 107, 117
Multiple correlation, *see* Correlation, multiple
Mumford, H. W., 136

N

National Bureau of Economic Research, 54
Net correlation, *see* Correlation, partial
Net regression, *see* Regression equations, linear; Regression equations, partial
Non-determination, coefficient of, 160
Normal equations, solution, 147, 174, 431
Normal frequency curves, 304
Null hypothesis, 319, 349

O

Ogburn, W. F., 164
Ogive, 9, 10
One-way tables, *see* Tabulation analysis
Origin, arbitrary, 18, 47, 155

P

Paired differences, 336, 337
Part correlation, 198
Partial correlation, *see* Correlation
Partial regression coefficient, 168
Pearson, F. A., 89, 102
Pearson, H., 138
Pearson, Karl, 28, 54
Percentage, of corresponding month, cycles, 116
  of preceding year, cycles, 106
Percentiles, 40
Persons, W. M., 98
Platykurtic, 54
Polygon, 9
Pooled standard deviation, 313
Population, 302, 304
  correlation from samples, 401
  distribution of samples from, 319
  standard deviation in means from, 308
  χ^2 compared to sample, 391
Price relatives, *see* Index numbers
Probable error compared with standard, 310
Product moment, 152, 169, 172
Product-moment method, 150, 152, 153, 155
Proportions, standard error of, 314
  standard error of difference between, 314
Purchasing power compared with deflated, 109
Purchasing-power method, 109, 117

Q

Quartiles, deviation, 38, 49
  first, 37
  range, 37
  semi-interquartile range, 37
  standard error of, 315
  standard error of difference between, 316
  third, 37

R

Random sample, 301
Range, 36, 49
  quartile, 37
  semi-interquartile, 37
Ratcliffe, H. E., 245
Reed, W. G., 162
Regression coefficients, 161, 168
Regression curves, 204, 213, 216, 220–222, 225–227, 229, 231–235
Regression equations, curvilinear, 202
  curvilinear joint, 250, 261
  curvilinear multiple, 212–215
  from tabular analysis, 289–291
  linear gross, 147, 160
  linear joint, 247, 292
  linear multiple, 176, 281
    uses, 183
  partial, 168
  significance of estimates from, 419
  tabular presentation, 164, 183, 251, 257
Relationships, *see* Correlation; Tabulation analysis
Relatives, *see* Index numbers
Reliability, measures of, 300
  of correlation analysis, 401
  of tabulation analysis, 323, 370, 387
Rho (ρ), *see* Correlation
Root-mean-square deviation, *see* Deviation, standard
Ruler method for linear trend, 76

S

Sample, distribution of, 319
  generalizing from, 307
  population correlation from, 401
  random, 301
Sample—(*Continued*)
  small, 320
  standard deviation in means from, 308
  χ^2 compared to population, 391
Sarle, C. F., 162
Schickele, R., 340
Seasonal variation, elimination of, 101
  methods of calculating, comparison, 97
    link-relative, 95
    moving-average, 93
    simple averages, 90
    trend-adjusted, 91
  test for, 365
  uses, 99
Secular trend, linear, methods of determining, averages, 78
    least-squares, 79
    ruler or string, 76
    selected-points, 78
    semi-average, 78
  non-linear, methods of determining, exponential curve, 83
    moving-average, 83
      calculation of, 83
      cutting corners, 87
  uses, 88
Selected-points method for linear trend, 78
Semi-average method for linear trend, 78
Semi-interquartile range, 37
Significance, *see* Reliability; Standard error; Analysis of variance; Chi square
Significant, defined, 325
Simultaneous equations, solution, 147, 174, 431
Skewness, 36, 54
Smith, B. B., 182
Snedecor, G. W., v, 350, 388, 412
Spencer, L., 102
Standard deviation, *see* Deviation, standard
Standard error, 304
  and probable errors, 310
  applied to tabular analysis, 323
    consistency of relationships, 327
    difference between two means, 325
    one-way frequency tables, 339
    paired differences, 336
    single mean, 323

- Standard error—(*Continued*)
  - applied to tabular analysis—(*Continued*)
    - two-way frequency tables, 341
    - two-way percentage tables, 343
    - two-way tables of averages, 329
      - diagnosis of, 333
  - compared with F and χ^2 tests, 399
  - normal frequency curves, 304
  - null hypothesis, 319, 349
  - of average deviation, 316
  - of correlations, gross, 405
      - z transformation, 407
    - partial, 408
      - z transformation, 409
  - of difference between two average deviations, 316
  - of difference between two correlations, 414
    - gross, 415
    - partial, 416
  - of difference between two frequencies, 314
  - of difference between two means, 311
  - of difference between two medians, 315
  - of difference between two proportions, 315
  - of difference between two quartiles, 316
  - of difference between two standard deviations, 316
  - of estimate, 167, 419
  - of estimates based on regression, 419
  - of frequencies, 314
  - of mean, 304–307
    - from population, 308
    - from sample, 308
    - of paired differences, 336
  - of median, 315
  - of proportions, 314
  - of quartiles, 315
  - of second differences, 313, 332
  - of standard deviation, 316
  - probability of occurrence of T, 319
  - probability of occurrence of t, 320
  - T, 316
  - t, 320
    - degrees of freedom for, 322
    - uses, 322
  - uses, 322
- Straight-line trend, *see* Secular trend
- String method of linear trend, 76

T

- T, 316
  - table of, 319
- t, 320
  - degrees of freedom for, 321
  - t test *vs.* F test, 368
  - table of, 320
  - uses, 322
- Table, of F, 350
  - of t, 320
  - of χ^2, 388
- Tabular method, *see* Tabulation analysis
- Tabulating machines, sums of products and squares with, 425
- Tabulation analysis, absence of relationships, 128
  - additive and joint relationships, 283
    - compared with correlation, 293
    - differences, first, second, and third, 284–287
    - rates of change, 288
  - additive relationships, 128, 283
  - advantages and disadvantages, 141, 266, 273, 276, 282, 293, 295–299
  - analysis of variance, applied to additive relationships, 374
    - applied to curvilinear relationships, 377
    - applied to differences between means, 354
      - t test *vs.* F test, 356
    - applied to joint relationships, 374
    - applied to non-numerical variables, 360
    - applied to one-way classification, 357
      - consistency of relationships, 358
    - applied to three-way tables, 381
    - applied to two-way classifications with equal subgroups, 360
      - more than one observation in subgroups, 360
      - one observation in subgroup, 365
    - applied to two-way table with unequal subgroups, 370

- Tabulation analysis—(*Continued*)
  - characteristics, 140
  - chi square applied, 389
  - curvilinear relationships, 127, 269
  - holding interrelated variables constant, 139
  - interserial relationships, 274
    - compared with correlation, 276
  - joint relationships, 132, 283
  - linear relationships, 126, 264
    - compared with correlation, 266
  - multiple relationships, four variables, 277
    - compared with correlation, 282
  - multiple relationships, three variables, 270
    - compared with correlation, 273
  - non-numerical, dependent variables, 137
    - independent variables, 134
      - one-way, 135
      - three-way, 136
      - two-way, 135
  - numerical variables, four-way, 125
    - higher-order, 125
    - one-way, 120
    - three-way, 124
    - two-way, 123
  - standard errors applied, 323
    - consistency of relationships, 327
    - difference between two means, 325
    - one-way frequency tables, 339
    - paired differences, 337
    - single mean, 323
    - two-way frequency tables, 341
    - two-way percentage tables, 343
    - two-way tables, 329
      - diagnosis of, 333
  - t and F tests compared, 368
  - t, F, and χ^2 tests compared, 400
- Tabulation *vs.* correlation, 264
  - flexibility of methods, 296
  - non-numerical variables, 297
    - results, 297
  - number of observations, 296
  - simplicity of methods, 295
- Thomsen, F. L., 103
- Time-reversal test, 70
- Timoshenko, V. P., 170
- Tolley, H. R., 162
- Trend, long-time, *see* Secular trend
- Tufts, W. P., 162
- Two-dimensional graph, *see* Correlation, joint

U

- Underwood, F. L., 132, 334
- Unequal subgroups, analysis of variance applied, 370
- Universe, 301
- Unweighted, *see* Index numbers

V

- Variability, coefficient of, from average deviation, 43
    - from quartile deviation, 40
    - from standard deviation, 49
    - uses, 52
  - importance, 36
  - measures of, 49
  - subdivision of, 345
  - unaccounted for, causes, 176
- Variables, dependent, defined, 122
  - holding interrelated constant, 139
  - independent, defined, 122
- Variance, 348
  - about a line and mean, 149
  - analysis of, *see* Analysis of variance
  - ratio, *see* F
- Vass, A. F., 138
- Vial, E. E., 181, 182

W

- Waite, W. C., 103
- Wallace, H. A., 162
- Warren, G. F., 89
- Weighted differences, 130
- Weighted indexes, 62
- Weights, determination of, 62, 71
  - variable, 72
- White, O. H., 135

Z

- z transformation, 407, 409
- Zero-order correlation, *see* Correlation, gross